UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS

Unraveling the genomics of adaptation of *Colletotrichum kahawae* to *Coffea arabica*

" Documento Definitivo"

Doutoramento em Biologia
Especialidade de Biologia Evolutiva

Ana Cristina Magalhães Vieira

Tese orientada por:
Doutora Dora Cristina Vicente Batista Lyon de Castro
Prof. Doutor Octávio Fernando de Sousa Salgueiro Godinho Paulo

Documento especialmente elaborado para a obtenção do grau de doutor

2018

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS

Unraveling the genomics of adaptation of *Colletotrichum kahawae* to *Coffea arabica*

Doutoramento em Biologia

Especialidade de Biologia Evolutiva

Ana Cristina Magalhães Vieira

Tese orientada por:

Doutora Dora Cristina Vicente Batista Lyon de Castro

Prof. Doutor Octávio Fernando de Sousa Salgueiro Godinho Paulo

Júri:

Presidente:

- Doutora Maria Manuela Gomes Coelho de Noronha Trancoso, Professora Catedrática e Presidente do Departamento de Biologia Animal da Faculdade de Ciências da Universidade de Lisboa

Vogais:

- Doutor José Paulo Sampaio, Professor Associado
  Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa;
- Doutor Artur Jorge da Costa Peixoto Alves, Investigador equiparado a Investigador Principal
  Departamento de Biologia da Universidade de Aveiro;
- Doutora Maria Helena Mendes da Costa Ferreira Correia de Oliveira, Professora Associada
  Instituto Superior de Agronomia da Universidade de Lisboa;
- Doutora Maria da Luz da Costa Pereira Mathias, Professora Catedrática
  Faculdade de Ciências da Universidade de Lisboa;
- Doutora Célia Maria Romba Rodrigues Miguel, Professora Auxiliar
  Faculdade de Ciências da Universidade de Lisboa;
- Doutora Dora Cristina Vicente Batista Lyon de Castro, Bolseira Pós-Doutoramento
  Faculdade de Ciências da Universidade de Lisboa (orientadora).

Documento especialmente elaborado para a obtenção do grau de doutor

Fundação para a Ciência e Tecnologia (SFRH/BD/89397/2012)

2018

Agradecimentos

O doutoramento, foi para mim, uma jornada de aprendizagem e auto descoberta indescritível, tanto a nível pessoal como a nível profissional, e a mesma não teria sido possível sem a ajuda de diversas pessoas às quais gostaria de agradecer.

Em primeiro lugar, gostaria de agradecer aos meus orientadores, Dora Batista e Octávio S. Paulo, por me acolherem na vossa equipa, me lançarem este desafio e me orientarem em todo o processo. A ti Dora, agradeço especialmente toda a motivação que me transmitiste durante estes anos. Sei que para ti as coisas nem sempre foram fáceis, mas independentemente disso, havia sempre um sorriso e uma palavra de conforto. A tua curiosidade cientifica é absolutamente contagiante, inspira-nos a querer saber sempre um pouco mais, enriqueceu os nossos debates científicos, e foi crucial para o meu amadurecimento cientifico. A si professor Octávio, gostaria de lhe agradecer pelos debates científicos que tivemos, que contribuíram para a interpretação dos resultados, e fizeram com que muitas vezes os conseguisse analisar de uma perspetiva diferente. Gostaria também de lhe agradecer toda a partilha de conhecimento, especialmente na área da biologia evolutiva, que foi vital para o meu amadurecimento cientifico.

Em segundo lugar, gostaria de agradecer aos meus amigos e colegas de trabalho que fazem ou fizeram parte do Cobig2 e do CIFC. Especificamente, gostaria de agradecer a todo o staff do CIFC (Sandra, Idalina, Miguel e Paula Leandro) por toda a ajuda na manutenção das coleções, montagem de ensaios, inoculações, preparação do material para observação microscópica, entre outros. Ao Engenheiro Vítor e a Dr. Céu pela partilha do vosso enorme conhecimento cientifico no modelo biológico. À Leonor, por ter partilhado comigo as suas “skills” na organização de trabalho e gestão de tempo. À Dr. Paula, por ter cuidado sempre tão bem das minhas “plantinhas”, a sua ajuda foi preciosa para a realização de todos os ensaios de agressividade. À Andreia e a Inês, pela amizade, e por toda a ajuda que me deram em diversos passos deste doutoramento. Ao Francisco, pela sua amizade, pelo enriquecimento do meu conhecimento “geek”, e por manter as condições necessárias para a realização de toda a análise bioinformática. Aos restantes membros do CIFC e do Cobig2, quero agradecer o ambiente acolhedor que me proporcionaram para a realização deste trabalho, e por todo o “feedback” positivo e ajuda que sempre me deram.

Ao Paulo Borges, por me ter dado a oportunidade de trabalhar num novo e excitante projeto, enquanto me fornecia todas as condições necessárias para terminar o meu doutoramento.

Por fim, tenho de agradecer às pessoas mais importantes da minha vida, a minha família e os meus amigos. Caetano, meu fiel companheiro da anormalidade, és sem duvida uma das

melhores pessoas que conheci até hoje. A tua amizade foi crucial para a manutenção da minha sanidade mental, não só pela quantidade de lixo da Internet que partilhaste comigo, mas principalmente por todas as nossas diarreias cerebrais. Fernando (aka pai, filho, Manel), a tua amizade é um dos meus grandes pilares e algo que eu não quero perder. Não existem palavras capazes de descrever o quanto te agradeço por tudo o que fizeste por mim durante este período, e apenas espero, um dia conseguir retribuir-te na mesma moeda. Inês, minha parceira de aventura, obrigada por me forçares a sair da tese e a ter um pouco de vida própria. Foi contigo que descobri uma das minhas atividades preferidas e visitei sítios fantásticos. Os nossos momentos de descontração foram cruciais para a recuperação de energia. Joaninha, minha pequena guerreira, és o meu exemplo de coragem, determinação e resiliência, e espero um dia conseguir lutar pelos meus sonhos com a mesma intensidade com que tu lutas pelos teus. Companheiros da jogatana, Bin, Simões, Freezer, André Reis e Rita, obrigado por todos os momentos de descontração, gargalhadas, convívio e amizade. Por me levarem para outra realidade onde posso partir, esmagar e atirar coisas, só porque sim. Por se esforçarem tanto para enriquecer o meu conhecimento “geek”, por me acolherem no vosso grupo, mesmo não gostando das mesmas coisas que vocês, e principalmente, por me mostrarem a beleza dos jogos de tabuleiro e D&D. Ao Tiago, Rafaela, Bruno e Ines, por todos os momentos de descontração e diversão, pelos jantares gourmet, pela gritaria fabulosa que muitas vezes me invade a casa e por todas as parvoíces que partilhamos nos últimos anos. O Mateus e a Madalena serão certamente crianças traumatizadas, mas muito felizes!

À minha família do sul, Avó, Mamy e Sr. Gustavo, que me acolheram no seu núcleo familiar como se fosse sua filha e neta. A vossa ajuda, apoio e disponibilidade constantes foram cruciais para a conclusão deste trabalho. Durante este período estiveram sempre atentos ás nossas necessidades. Oferecendo-nos ajuda, mesmo antes de sabermos que precisávamos dela, e acreditando sempre em nós. Essa confiança “cega” na qualidade e importância do nosso trabalho foram vitais para a nossa auto-motivação, e sem isso, por vezes tudo teria sido mais difícil.

À minha família por serem, apesar da distancia, uma das minhas principais fontes de energia e motivação. Por terem acreditado sempre em mim, mesmo quando eu teimava em não acreditar, e especialmente, por muitas vezes me terem dado o empurrão necessário para que não desistisse. Aos meus pais, Mamy e Cotinha, por me terem proporcionado todas as condições necessárias para concretizar os meus sonhos. Sei que nem sempre foi fácil e que se questionaram muitas vezes se tomaram a decisão certa, mas espero que saibam, o quão grata vos sou por me terem dado esta oportunidade. Aos meus irmãos e cunhados, por serem um pilar inabalável na minha vida, por estarem sempre disponíveis e há distancia de um telefonema. Por todas as vezes que me deram a mão e me ajudaram a levantar, por todas as

vezes que me ouviram, me limparam as lágrimas ou simplesmente festejaram comigo as minhas conquistas. Aos meus sobrinhos e afilhado, por cada sorriso, abraço e beijinho, por todas as vezes que gritaram o meu nome com entusiasmo, ver-vos crescer tem sido uma das experiências mais gratificantes da minha vida, e muitas vezes, o vosso carinho iluminou o dia mais nublado.

Por ultimo ao Diogo, o amor da minha vida, por tudo aquilo que partilhamos. Entramos nesta aventura sem saber o que esperar, mas certos que o enfrentaríamos juntos, e portanto, falhar não era uma opção. O doutoramento foi um dos períodos mais felizes e difíceis da minha vida, e se por um lado, partilhávamos uma vida em conjunto com toda a nossa cumplicidade, por outro, tivemos de superar grandes adversidades. Nem sempre foi fácil, mas caminhamos juntos, de mãos dadas, com a certeza de que nunca cairíamos porque não estávamos sozinhos, e hoje estamos aqui, prestes a terminar. És o meu melhor amigo, o meu companheiro de viagem, a minha maior fonte de motivação e auto confiança, e a pessoa com a qual partilho tudo, e portanto, a ti te devo o meu maior obrigado. Espero que saibas o quanto gosto de ti, o quanto valorizo a nossa relação, e principalmente a falta que me fazes quando estás longe. Amo-te, mais do que tudo! Até já!!

Table of Contents

Abstract

Fungi are one of the most destructive groups of plant pathogens, being responsible for more than 30 % of the emerging diseases worldwide. *Colletotrichum kahawae* is the causal agent of Coffee berry disease, one of the most devastating diseases of *Coffea arabica* in Africa. This pathogen is able to specifically infect green coffee berries, leading to severe production losses if no control measures are applied. In this thesis, we used an integrative approach, joining population genomics, pathology and gene expression, to better understand several aspects of this pathogen's population dynamics and adaptation, including its evolutionary history, pathogenicity and aggressiveness. The population genomics analysis, revealed that *C. kahawae* is a true clonal pathogen, that probably emerged in Angola, quickly dispersed to East Africa, and only later colonized the Cameroon. Two clonal lineages within the Angolan population were detected and the evolutionary mechanism that gave rise to the Cameroonian population was not evident, leading to the proposal of alternative hypotheses. High genomic differentiation between *C. kahawae* and closely related species was detected, reinforcing its recognition as a distinct species. Further RAD-seq analysis uncovered the genomic loci putatively underlying *C. kahawae's* pathogenicity. The annotation of these loci, suggest that several biological processes (detoxification, transport, signaling, and regulation of host and pathogen gene expression) could be involved in host infection. For testing association with aggressiveness, *C. kahawae* phenotypic profiling was performed and three different aggressiveness classes were created to accommodate all the variability observed. The genome-wide association analyses performed were able to detect a group of SNPs of small effect, providing candidate genes putatively associated with aggressiveness. Finally, aiming a follow-up validation of the functional role of candidate genes, the first steps to perform gene expression studies were followed. The best strategy to normalize a broad range of fungal and interaction samples in isolates with different aggressiveness profiles was established.

Keywords: Plant pathogen; host-pathogen interaction; population genomics; gene-expression; pathological tests

Resumo

Atualmente, os fungos fitopatogénicos são responsáveis por aproximadamente 30% das perdas na produção agrícola mundial e constituem um grave problema de segurança alimentar. Além disso, nos últimos anos temos assistido a um aumento significativo do número de novos e emergentes fungos fitopatogénicos, e portanto, torna-se crucial o melhoramento das estratégias de proteção das plantas. Assim sendo, apenas uma abordagem multidisciplinar será capaz de efetivamente contribuir para uma agricultura sustentável, e portanto, informações de várias áreas da ciência, nomeadamente da biologia evolutiva, devem ser tidas em conta para alcançarmos o sucesso.

Neste contexto, é fundamental compreender melhor a história demográfica de um agente patogénico emergente, assim como o seu potencial evolutivo e capacidade de superar as barreiras de defesa da planta. De facto, sabe-se que as plantas e o seu agente patogénico travam uma batalha evolutiva contínua, na qual os agentes patogénicos tentam suprimir as defesas da planta, e esta por sua vez, tenta prevenir ou mitigar os danos provocados pelo agente patogénico. Num sistema natural, as força dos dois intervenientes estão equilibradas, mas no sistema de monocultura, o agente patogénico tem uma vantagem seletiva, devido à elevada homogeneidade genética do hospedeiro, que potencia a sua dispersão após a quebra da barreira de resistência da planta.

A antracnose dos frutos verdes é uma das principais doenças do cafeeiro Arabica, atualmente restrita ao continente africano, a qual pode levar a perdas de produção na ordem dos 80%, se nenhuma medida de controlo for aplicada. O seu agente causal, *Colletotrichum kahawae* Waller & Bridge, encontra-se perfeitamente adaptado aos frutos verdes do cafeeiro, e pensa-se que emergiu através de um salto de hospedeiro de uma espécie não patogénica e geneticamente muito próxima. Este agente patogénico é considerado como uma arma biológica e a sua potencial dispersão para outras regiões produtoras de café na Ásia e América central é muito temida. Até hoje, nenhuma medida de controlo totalmente eficaz foi implementada, e portanto, torna-se crucial melhor compreender a dinâmica populacional e evolução adaptativa deste patógeno, de forma a contribuir para a implementação de medidas de controlo mais eficientes e sustentáveis. Por conseguinte, na presente tese foi utilizada uma

abordagem integrativa, conjugando genómica, patologia e expressão génica para investigar e melhor compreender processos adaptativos, incluindo patogenicidade e agressividade, e a historia evolutiva de *C. kahawae*.

A análise de um painel de milhares de polimorfismos nucleotídicos únicos (SNPs), obtidos pela técnica de "RAD-seq", permitiu concluir que *C. kahawae* é um agente patogénico puramente clonal, com baixa capacidade de dispersão e potencial evolutivo, que provavelmente terá emergido em Angola. Três populações geneticamente distintas (Angola, Camarões e África Oriental) foram identificadas, entre as quais a população de Angola e de África Oriental, parecem ter surgido praticamente ao mesmo tempo, e só mais tarde a população dos Camarões terá emergido a partir da população de Angola. Além disso, duas linhagens clonais foram encontradas na população Angolana, e o mecanismo evolutivo que terá dado origem à formação da população dos Camarões não foi completamente descortinado, propondo-se hipóteses alternativas. Na verdade, permanece por esclarecer se *C. kahawae* é um agente patogénico puramente clonal, ou se na presença de condições muito hostis é capaz de recombinar e se adaptar mais rapidamente. Finalmente, os nossos resultados sugerem que se forem cumpridas as medidas fito sanitárias já existentes a probabilidade de dispersão deste agente patogénico para outras regiões produtoras de café é bastante reduzida.

Adicionalmente, foi aplicada uma estratégia de genómica comparativa com recurso a "RAD-seq" entre *C. kahawae* e três espécies próximas não patogénicas, com o objetivo de identificar os mecanismos genéticos subjacentes à sua patogenicidade, isto é a capacidade de infetar os frutos verdes do cafeeiro. Neste trabalho foi detetado, pela primeira vez, uma enorme diferenciação genética entre *C. kahawae* e as espécies próximas, o que reforça a ideia de que este agente patogénico constitui na verdade uma espécie distinta. Além disso, foi também identificado um grande número de SNPs de diagnóstico (5 560 SNPs), isto é conservados entre todos os isolados patógenicos mas diferenciados de pelo menos um isolado não patogénico, havendo um grande enriquecimento do número de mutações não sinónimas neste conjunto de dados. A anotação dos genes com um enriquecimento no número de mutações funcionais, e consequentemente, potencialmente envolvidos na patogenicidade, evidenciou que este é um mecanismo de infeção altamente complexo e dependente de vários processos biológicos, tais como a desintoxicação e transporte, regulação das respostas do hospedeiro e sinalização.

Por sua vez, a identificação das regiões genómicas potencialmente envolvidas na agressividade, requereu a caracterização fenotípica do perfil de agressividade dos isolados em estudo. Para tal, foi utilizado um vasto conjunto de métricas, possibilitando o estabelecimento de três classes de agressividade capazes de acomodar toda a variação fenotípica observada. Indivíduos representativos de cada uma dessas classes foram analisados microscopicamente, o que permitiu concluir que a agressividade está relacionada com o desenvolvimento de estágios de pós-penetração, ao invés de germinação de conídios e desenvolvimento de apressórios. O estudo de associação genómica, para testar a relação dos SNPs com a característica da agressividade, não permitiu a identificação de SNPs causais, mas permitiu a identificação de SNPs de pequeno efeito (15 na análise de multi-associação, 10 na análise de associação única e 7 comuns entre as duas abordagens), que embora não sejam capazes de explicar toda a diversidade fenotípica observada, fornecem um conjunto de genes candidatos putativamente associados a esta característica. Os resultados obtidos podem ter sido limitados pela verdadeira clonalidade de *C. kahawae,* que dificultou a identificação integral dos SNPs associados, ou podem mesmo sugerir que a agressividade é essencialmente regulada por mecanismos de expressão diferencial e não tanto por variações alélicas. Por outro lado, a agressividade é uma característica fisiológica muito suscetível a diversas condições ambientais (estado fisiológico do hospedeiro e agente patogénico, condições de cultura, temperatura), e por conseguinte, pode não estar sujeita a pressões da seleção natural. No entanto, um conjunto de genes candidatos, potencialmente associados a ambos os processos adaptativos (patogenicidade e agressividade) foi identificado, e o seu papel na manutenção desta complexa interação deve ser investigado através de estudos de expressão génica e de desenvolvimento de mutantes.

Por fim, foram iniciados estudos de expressão génica através do estabelecimento da melhor estratégia de normalização para efetuar a calibração na análise da expressão, tendo por base isolados com diferentes perfis de agressividade e um alargado conjunto de amostras referentes a diferentes estádios de desenvolvimento do fungo e da interação com o hospedeiro. De um modo geral, o perfil de agressividade dos isolados não influenciou a estratégia de normalização aplicada, mas sim o tipo de amostra (fungo e interação), e consequentemente, um conjunto diferente de genes têm de ser utilizado para normalizar as diferentes amostras. Este estudo foi crucial para fornecer

as ferramentas necessárias à implementação de estudos funcionais futuros neste âmbito.

Em suma, os resultados obtidos neste trabalho demonstram o poder de se aplicar uma abordagem integrativa no estudo de uma doença emergente. Concretamente, permitiu-nos perceber melhor a dinâmica populacional, história e potencial evolutivo e capacidade de dispersão deste agente patogénico, assim como os mecanismos genéticos associados à sua patogenicidade e agressividade. Contudo, todos estes processos biológicos estão longe de estarem completamente compreendidos, e portanto, um conjunto de novos desafios foram lançados no decorrer desta tese, que espero, que sejam futuramente respondidos.

Palavras-chave: interação planta-agente patogénico; fungos fitopatogénicos; expressão genica; estudos de patologia; genética populacional

Figure Index

Index of Tables

General Introduction 1

1.1 Emergent plant pathogens: a worldwide problem

1.1.1 The adaptive capacity of fungal plant pathogens

In the past few years, there has been an unprecedented number of new and emerging pathogenic fungi able to cause severe crop diseases and production losses, which represent one of the greatest problems ever for sustainable agricultural production and global food security. Some examples in the past show that pathogen emergence and subsequent epidemics are even able to change the course of human history. For instance, in the late potato blight that led to starvation, economic ruin and the downfall of the English government during the Irish potato famine, and in the twentieth century, when Dutch elm blight and chestnut blight destroyed urban and forest landscapes (Fisher *et al.*, 2012). Nonetheless, one of the most well known and impressive episodes in history of devastating social and economic consequences brought by a plant disease epidemics was the obliteration of coffee cultivation from Ceylon (Sri Lanka) by Coffee Leaf Rust, which led to a switch in agricultural production from coffee to tea (McCook and Vandermeer, 2015).

To face this problem it has become evident that only a multidisciplinary approach, in which the ecoevolutionary principles are taken into account to slow down the rates of evolution of crop pathogens, will work on a long-term scenario (McDonald and Stukenbrock, 2016; Pautasso *et al.*, 2012; Zhang *et al.*, 2015). The timescales involved in pathogen emergence range from thousands of years for pathogens that emerged through co-evolution of host and pathogen, to practically instantaneous for new pathogen races or species that emerged through hosts jumps, host domestication, clonal divergence, horizontal gene transfer or inter-specific hybridization (Giraud *et al*., 2010).

In this sense, the use of genomic tools in evolutionary biology will allow the reconstruction of the pathogen's evolutionary history and the processes that caused their emergence, including migration routes and population structures of ancestral and contemporary populations, as such processes affect the distribution and evolution of diversity (Barrés *et al.*, 2012; Estoup and Guillemaud, 2010). Moreover, it is also important to determine whether the adaptive evolution occurs within the invasion time range or whether the successful invaders were already well suited for establishment and spread before their introduction, and which traits or situations facilitate the invasion (Gladieux *et al.*, 2015). Many invasive fungal populations are under sustained pressure to adapt to environments distinct from their endemic ranges. The response to these novel environments can lead to invasive fungi emerging on new hosts, colonizing new varieties of their hosts, or re-emerging with greater pathogenicity (Gladieux *et al.*, 2015).

In fact, pathogenic fungi have a group of life-history features able to promote their evolutionary potential and rapid ecological speciation, namely: i) the high number of spores produced, as they increase both the possibility of survival on a new host and the levels of adaptive variation created by mutations; ii) mating within host, creating pleiotropy between host adaptation and assortative mating; iii) the strong disruptive selection imposed by the host; iv) frequent asexual reproduction with rare events of sexual recombination (Giraud *et al.*, 2010a; McDonald and Linde, 2002). The asexual reproduction facilitates speciation processes, as this correspond to multiple cycles of selection for local adaptation without recombination breaking down locally advantageous allelic combinations and introducing locally deleterious immigrant alleles (Haag *et al.*, 2013; Weir *et al.*, 2016). The mixture of sexual and asexual reproduction present in most fungal pathogens allows both the creation of new genetic combinations and the rapid amplification by selection of those combinations that promote infection of a new host (Bazin *et al.*, 2014; Dutech *et al.*, 2012; Dutech *et al.*, 2017; Weir *et al.*, 2016). Consequently, pathogens that pose the greatest risk of breaking down resistance genes have a mixed reproduction system, a high potential for genotype flow, large effective population sizes, and high mutation rates. The lowest risk pathogens are those with strict asexual reproduction, low potential for gene flow, small effective population sizes, and low mutation rates (McDonald and Linde, 2002). Finally, the unifying driving force for these different evolutionary processes is the nature of the agro-ecosystems,

and consequently, the boom-and-bust cycles of disease epidemics result from complex interactions among hosts, pathogens, and environments (classic disease triangle), that drive changes in population genetics composition and pathogen evolutionary trajectories (Grünwald *et al.*, 2016; Pautasso *et al.*, 2012).

1.1.2 The evolutionary battle of plants and their pathogens

Plants and pathogens have engaged in a continuous evolutionary battle, with pathogens attempting to circumvent plant defense mechanism and plants responding through enhanced immune systems to prevent or mitigate the damage induced by the pathogen attack (Zhan *et al.*, 2014; Zhan *et al.*, 2015). In addition, plant-pathogen interaction is significantly affected by the environment, and therefore, some human activities (agronomic practices, fungicide treatments, movement of plant material in the global market) have an overwhelming impact in the co-evolution dynamics of disease epidemics (Elad and Pertot, 2014).

In natural ecosystems, the antagonistic interaction between plants and their pathogens shape the genetic variation at the genomic and population levels and leads to an evolutionary equilibrium. Pathogens usually have a selective advantage, due to their short generation time, large population size and high mutation rates, all of which enhance their ability to quickly respond to changes in host defenses. Meanwhile, the genetic variation and environmental heterogeneity among the host population, significantly brings down the infection capacity of the pathogens, reducing their ability to find the right host and right time and place to infect (Burdon *et al.*, 2013; Möller and Stukenbrock, 2017; Zhan *et al.*, 2014; Zhan *et al.*, 2015).

This concordant pattern of host-pathogen co-evolution is disrupted by agricultural practices in favor of pathogens (Zhan *et al.*, 2014). Agricultural ecosystems are composed of densely and genetically uniform crop populations, which have been selected for a targeted trait through crop breeding. Moreover, very few changes in regional crop species composition are observed from year to year, and the global trade in agricultural products and technologies make the planted fields remarkably similar across the continents. This standardization of host populations favors the emergence of new, host-specialized, “domesticated” crop pathogens that evolve more rapidly and are

more virulent than their “wild” ancestors. This evolutionary advantage also means that a strain able to infect one individual can quickly spread over an entire field or region, increasing the population size of the pathogen. Additionally, the global movement of pathogens (by passive vectors able to carry viable spores, or through the movement of infected material) promotes the emergence of new hybrid pathogens by mixing native and introduced pathogen species, which deeply increase their evolutionary potential (McDonald and Stukenbrock, 2016; Möller and Stukenbrock, 2017; Zhan *et al.*, 2014; Zhan *et al.*, 2015).

In addition, within this system, resistance genes in the hosts are one of the strongest drivers of pathogen evolution, as a successful infection is required for reproduction and dispersal, and therefore pathogenicity related genes are under strong selective pressure in pathogen genomes. Thus, despite the considerable investment required to develop new resistant varieties or agrochemicals, their effectiveness may be significantly reduced within only a few years of deployment (Zhan *et al.*, 2015). In fact, a single and static disease control management is bound to fail sooner or later regardless of how effective the approach may be at its inception (Zhan *et al.*, 2014). Therefore, our efforts should be taken to reduce the amplitude of disease epidemics in the short term, and the evolutionary potential of pathogens in the long term, without imposing strong selection on the pathogen or a negative effect on the environment. In essence, sustainable disease control should be achieved through an integrative disease management program that is able to minimize the evolutionary potential of plant pathogens by reducing their genetic variation, stabilize their evolutionary dynamics by diversifying selection, and restrict the migration of pathogens that carry new virulence genes (McDonald and Stukenbrock, 2016; Zhan *et al.*, 2014; Zhan *et al.*, 2015). In this sense, the current thesis was developed to better understand the population dynamics, the host adaptation and evolutionary potential of an important coffee pathogen, through an integrative approach.

1.2 The Coffee crop

Coffee is one of the most important commodities in the international agricultural trade (**Figure 1.1**), representing a significant source of income in many tropical and

subtropical regions across Latin America, Africa and Asia (ICO 2017). It provides livelihood for over 125 million people worldwide through its growing, processing and trade, and in 2017/2018 accounted for exports worth of an estimated value of US$ 20 billion (ICO 2017). Globally, only two coffee species are used in commercial production: *Coffea arabica* L. and *Coffea canephora* Pierre ex Froehner, also known as Robusta coffee. *C. arabica* is adapted to cool and humid environmental conditions at higher altitudes (1300-1800m) and is quite susceptible to several diseases (Silva *et al.*, 2006). This crop is considered the most important commercial species, representing 64% of the world coffee production (ICO 2017), and produces the best quality coffee with lower caffeine content (Hindorf and Omondi, 2011). By contrast, *C. canephora,* accounts for most of the remaining coffee trade, and is usually found in humid and warmer environments of the lowlands and seems to be more resilient to pathogen diseases (Silva *et al.*, 2006). In general, the genetic variability within coffee plantations is very low. In fact, it is believed that only two botanical varieties (*C. arabica* "Typica" Cramer and *C. arabica* "Bourbon" Choussy) gave rise to the vast majority of *C. arabica's* world plantations (Lécolier *et al.*, 2009). *C. canephora,* on the other hand, it was not cultivated until the beginning of the 20th century and its production is still restricted to the low lands (Bigger, 2006).

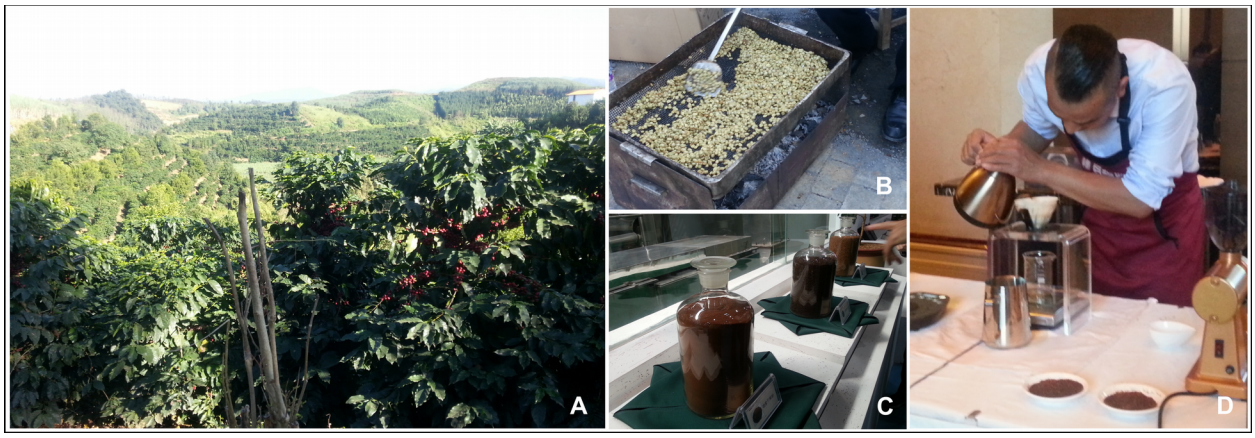


Figure 1.1 - An overview of Coffee production. **A)** Coffee plantations in China; **B)** Seeds drying – traditional method; **C)** Ground coffee**; D)** Chinese traditional coffee confection;

In fact, the domestication of coffee, as in other agronomic crops, was characterized by intensive artificial selection and severe bottlenecks, that deeply altered the populations of the wild progenitor species, into the domesticated varieties we know today (Anthony

et al., 2002). Consequently, this crop is susceptible to many pre and post harvest diseases caused by fungi, bacteria and virus. Within these diseases, Coffee berry disease (CBD) and Coffee leaf rust (CLR) have a prominence for their incidence and economic impact (Silva *et al.*, 2006; Talhinhas *et al.*, 2017), representing the greatest threat for coffee sustainable production.

1.2.1 Coffee berry disease

1.2.1.1 The disease

Coffee Berry Disease is an emergent and severe disease of Arabica coffee crops, caused by the fungus *Colletotrichum kahawae* Waller & Bridge (Waller *et al*., 1993). It is considered the main limiting factor of Arabica coffee production in the African continent, where it is endemic, leading to severe losses (70% to 80%) if no control measures are applied (Hindorf and Omondi, 2011; Motteram *et al.*, 2011; Silva *et al.*, 2006). In Ethiopia, for instance, one of the largest coffee growing countries in Africa, the national average losses due to CBD are estimated to range from 24 to 30%, and those may reach 90% during favorable seasons in some areas. If these percentages are translated into monetary value, Ethiopia suffers losses of 84 million US$ annually only due to average CBD outbreaks. These losses become even worse if we take into account that over 15 million Ethiopians depend directly or indirectly on the coffee industry (Alemu *et al.*, 2017; Giddisa, 2016; Hindorf and Omondi, 2011). Moreover, it is estimated that CBD along with chemical control costs, lead to the loss of US$ 300–500 millions in Arabica coffee production (Van Der Vossen, 2009), thereby reducing the competitiveness of these regions on the coffee market (Pinard *et al.*, 2012).

Although, *C. kakawae* is able to attack all stages of the developing plant including flowers buds, leaves, fruits and maturing bark, the disease occurs as a result of the infection of green coffee berries (Silva *et al.*, 2006). Symptoms on green berries appear as small dark sunken lesions, anthracnose-like lesions, which grow and eventually cover the whole berry, leading to mummification and premature dropping (**Figure 1.2**) (Giddisa, 2016). The mummified berries lose all commercial value and may be shed off the tree or remain attached as a source of inoculum (Pinard *et al.*, 2012). The sporulation of the pathogen is higher at the onset of rain events and it mainly occurs on

the surface of infected berries, leading to the creation of mucilaginous masses able to protect spores from dissection and loss of viability in dry weather (Silva *et al.*, 2006). The inoculum is dispersed by rain through “splash” in a short distance, or by passive vectors such as man, vehicles, birds, insects and infected plant material that may carry viable spores at long distances (Alemu *et al.*, 2017; Giddisa, 2016; Hindorf and Omondi, 2011). The occurrence and intensity of CBD varies from place to place and between seasons, depending largely on host susceptibility, pathogen aggressiveness and favorable weather conditions, such as low temperatures accompanied by high rainfall periods (Giddisa, 2016).

Figure 1.2 - Coffee Berry Disease in the field. **A**) Uganda coffee plantation, in which all the fruits from the tree are completely destroyed by the disease; **B**) Close-up of a tree, in which the green berries has different levels of infection

Several strategies to minimize production losses have been implemented, namely chemical control, host resistance and cultural practices (Bedimo *et al.*, 2007). Chemical control was the first and the most popular attempt to manage the disease outbreaks. Unfortunately, this strategy, revealed itself to be somewhat inefficient, whether because chemicals are washed away in the rainy seasons, when the fungus strikes most, or because the fungus rapidly acquires tolerance in the field (van den Bosch and Gilligan, 2008; Chung *et al.*, 2006). Moreover, CBD is endemic to Africa where most smallholders are unable to carry out the spray programs due to the lack of subsidies for fungicide application and sprayer purchases (Giddisa, 2016). Beyond that, analysis of coffee production costs revealed that chemical control of CBD alone contributed to up to 30% of the total (Derso and Waller, 2003), and it is impossible to bear this economic burden

in a long term scenario. Finally, a global trend to minimize the application of fungicides and consume only organic coffee has grown in popularity worldwide, increasing the economic advantage of “biological plantations” (Giddisa, 2016).

Coffee breeding programs have been a viable option within this context and were created to introduce resistance genes in commercial varieties that already had desirable traits, especially yield, quality and adaptability to coffee growing conditions (Gichuru *et al.* , 2008). Some countries including Ethiopia, Kenya, Tanzania have released resistant varieties for commercial production but the challenges of long breeding cycle associated with long juvenile period have slowed down the pace of further varietal improvement. Screening for varieties resistant to CBD in the field is difficult, as it depends on the occurrence of sustainable climatic conditions and disease expression (Pinard *et al.*, 2012), and consequently, the screening procedures still largely rely on controlled inoculations (van der Vossen *et al.*, 2015). Additionally, a significant effort has been made in Latin America, especially in Colombia, on the implementation of preventive selection strategies (Alkimim *et al.*, 2017; Pinard *et al.*, 2012; Silva *et al.*, 2006). Screening for resistance have been conducted in collaboration with the Coffee Rust Research Center (CIFC) in Portugal. This research center was created in 1955 to support the coffee breeding programs on the development of resistant varieties to *Hemileia vastatrix* (coffee leaf rust), and in 1989, it expanded its research for CBD resistance (Silva *et al.*, 2006). Consequently, today, CIFC has a collection of isolates from all different geographic locations where the disease exists, representative of the global *C. kahawae* range of aggressiveness, which can contribute to reliable pre-screening resistance tests within breeding programs. Nevertheless, breeding programs will be more efficiently together with improved agronomical practices. In fact, there are reports that some agronomic practices such as good aeration, rapid adequate pruning and wide spacing, shade control, artificial irrigation during the drying season, removal of mummified berries and inter-planted coffee trees with fruits trees, can create environmental conditions limiting CBD development (Garedew *et al.*, 2017; Giddisa, 2016). All this knowledge, associated to a better understanding of the population dynamics and evolutionary history of *C. kahawae,* can contribute for the design of sustainable and integrative control measures, able to significantly reduce the incidence of this pathogen in the field.

1.2.1.2 The pathogen – *Colletotrichum kahawae*

1.2.1.2.1 Biosecurity significance

Colletotrichum kahawae is a highly aggressive and specialized pathogen that belongs to the genus *Colletotrichum.* This genus has more than 100 species, which are considered major plant pathogens worldwide, as they cause significant economic damages to countless crops in tropical, sub-tropical, and tempered regions (Crouch *et al.*, 2014). In fact, due to its pervasiveness, capacity of destruction and scientific importance as model pathosystems, *Colletotrichum* spp. fungi are collectively ranked by the international plant pathology community among the top ten most important fungal phytopathogens (Crouch *et al.*, 2014). *C. kahawae*, specifically, is listed as a quarantine pathogen in Asia and Latin American and Australia (Batista *et al.*, 2017; Kebati *et al.*, 2016; Tao *et al.*, 2013), and as a biological weapon (Australia Group, 2014). Consequently, the pathogen's potential dispersal to other Arabica coffee cultivation regions outside Africa, particularly to those at high altitude is greatly feared (Batista *et al.*, 2017).

1.2.1.2.2 Taxonomy

The nomenclature of *C. kahawae* went through a long period of confusion until 1993 and nowadays went back to be a polemic topic (Batista *et al.*, 2017). The first *Colletotrichum* spp isolates collected from coffee plants were grouped according to their morphological traits in four groups: ccm (*C. coffeanum mycelial*), cca (*C. coffeanum acervuli*), ccp (*C. coffeanum* pink) and the CBD strain (Gibbs, 1969). The three former groups were latter recognized as *C. gloeosporioides* (ccm and cca) and *C. acutatum* (ccp), and proved to be non-pathogenic in green coffee berries, while the fourth group that was able to infect this niche was named *C. coffeanum* (Hindorf, 1970). Interestingly, the name *C. coffeanum* was not based on type material associated with CBD, but on samples from Brazil where the disease does not exist, and it clearly refers to *C. gloeosporioides* (Waller *et al.*, 1993). Thus, a significant effort was made to solve this nomenclature problem and in 1993 Waller and Brigde, based on several morphological, pathological and biochemical traits, finally described *C. kahawae* as the causal agent of CBD. Recently, this recognized species was brought down to a subspecies level (*C. kahawae* subsp. k*ahawae*), due to its remarkable genetic similarity with a generalist and

cosmopolitan group of *Colletotrichum* isolates, unable to infect the green coffee berries (*C. kahawae* subsp. c*iggaro*) (Weir *et al.*, 2012). Since then, there was a widespread confusion on the literature, with *C. kahawae* being reported all across the world and in different hosts (Batista *et al.*, 2017). Although these reports are referring to *C. kahawae* subsp. c*iggaro,* some of them could not distinguish the pathogen at the sub-specific level (Batista *et al.*, 2017). Currently, there is a state of fear and insecurity in the coffee growing community towards the potential dispersion of this harmful pathogen out of Africa, and all of this arose due to a taxonomic problem. Given the extreme impact that this situation may trigger and the subsequent biosecurity implications, there is a practical need to completely distinguish these pathogens taxonomically as to avoid the risk of misidentification (Batista *et al.*, 2017). Therefore, in this work, *C. kahawae* will be considered as an individual species being the taxonomy nomenclature adopted the following one:

Kindgom: Fungi

Phylum: Ascomycota

Sub-phylum: Pezizomycotina

Class: Sordariomycetes

Sub-class: Hypocreomycetidae

Order: Glomerellales

Family: Glomerellaceae

Genus: *Colletotrichum*

Species: C. kahawae

1.2.1.2.3 Origin and distribution

The origin of *C. kahawae,* which has been established largely based on historical data, remains difficult to pin point. Nevertheless, CBD was firstly reported in 1922 in Kenya, west of the Rift Valley, where it led to the decimation of several coffee plantations. In fact, this pathogen brought coffee cultivation in the west of the Riff Valley to a near end and tea plantations became predominant in the region (Giddisa, 2016; Hindorf and

Omondi, 2011). Despite the little attention received during the early stages of its documented emergence, African coffee growers soon wittiness a swift dissemination of CBD throughout most of the continent (Nutman and Roberts, 1960). Apparently, the free movement of coffee plant material from CBD infected areas to the remaining Arabica coffee-growing regions and the poor management practices, were the main dispersion factors of this pathogen, while environmental factors, except low altitude, did not make much difference in its dispersion (Giddisa, 2016). According to the historical data, CBD was further reported in Angola around 1930, Zaire in 1937, Cameroon in 1955-1957, Uganda in 1959, Tanzania in 1964 and Ethiopia in 1971 and in 1985 its presence was confirmed in Malawi, Zimbabwe and Zambia (Giddisa, 2016). However, a population genetic study showed that *C. kahawae* probably emerged in Angola, from a generalist group, unable to infect the green coffee berries [similar to the one described by Weir *at al.* (2012) as *C. kahawae* subsp. c*iggaro*] by a host-jump, and only after that, for the remaining coffee growing countries. The *C. kahawae's* diversification was estimated to have began at $<2200_{BP}$, leaving a very short time frame since the divergence from its sibling species (5600_{BP}), during which a severe drop in pathogen effective population size occurred without significant evidence of recovery until today. This results supports a scenario of recent introduction with a severe bottleneck followed by a subsequent adaptation to an unoccupied niche, the green arabica coffee berries. In such scenario, two intrinsic barriers to gene flow could have occurred, the immigrant inviability and the predominantly asexual behavior of *C. kahawae*. Both of which could have had an instrumental role in driving ecological speciation by creating pleiotropic interactions between local adaptation and reproductive patterns (Silva *et al.,* 2012).

1.2.1.2.4 Genetic and pathological variation

Several molecular markers, including multi-locus sequences, isoenzymes, RADPs and AFLPs, have been used to assess the genetic variation of *C. kahawae,* and low levels of genetic polymorphism were observed (Bridge *et al.*, 2008; Derso and Waller, 2003; Loureiro *et al.*, 2011; Luzolo *et al.*, 2010; Silva *et al.*, 2012). Brigde *et al.* (2008) suggested the existence of geographically differentiated populations based on the divergent AFLP banding pattern of three isolates from Cameroon and Malawi, from the remaining sampling. Manuel *et al.* (2010), using the ITS sequence region, also found

slight differences among isolates from Angola and between isolates from Angola and other east African countries. But only recently, Silva *et al.* (2012), using a multi-locus approach were able to detect three clonal and divergent populations: Angolan, Cameroonian and East African. The demographic pattern observed suggests a dispersion of the pathogen from Angola to Cameroon and from there to East Africa, with each population isolated in their respective highland areas. Moreover, significant differences among *C. kahawae* isolates' aggressiveness were early detected and have been referred in the literature (Beynon *et al.*, 1995; Loureiro *et al.*, 2011; Manga *et al.*, 1997; Pires *et al.*, 2016; Várzea *et al.*, 1999). However, a comprehensive characterization of *C. kahawae's* aggressiveness, using a large set of isolates and metrics, has not yet been performed and is still required. Despite that, *C. kahawae's* aggressiveness could be discriminated by an alkaline phosphatase isoenzyme (Loureiro *et al.*, 2011), and it was suggested to have a positive relationship with the number of mini-chromosomes (Pires *et al.*, 2016). However, no correlation has been found between genetic/molecular markers (e.g. RFLPs and RAPDs) and this trait (Beynon *et al.*, 1995; Derso and Waller, 2003).

1.2.1.2.5 Life-style and infection process

Concerning its life style, *C. kahawae* is a hemibiotrophic pathogen, as it exhibits a biotrophic phase followed by a necrotrophic phase. During the biotrophic phase, the pathogen is able to invade the host cells without killing them and feeds on the living tissues. Subsequently, the pathogen switches to a necrotrophic mode of nutrition, killing the host cells and feeding on the dead tissues (Silva *et al.*, 2006 and references within). Overall, the cycle of infection begins with the germination of conidia and differentiation of melanized appressoria, followed by host penetration (**Figure 1.3**).

In a susceptible plant, after penetration, the infection peg swells to form an infection vesicle inside the cell lumen, and further grows to intra and inter-cellular ramifications within the living host, which eventually leads to the switch between phases (Loureiro *et al.*, 2012; Silva *et al.*, 2006). The biotrophic phase is characterized by the absence of macroscopic symptoms and the intracellular infection vesicles and hyphae remain external to the plant plasma membrane. While the necrotrophic phase is characterized by the appearance of the first symptoms and is associated with severe cell wall

alterations and death of the host protoplast (Loureiro *et al.*, 2012; Silva *et al.*, 2006). The time required to activate the necrotic phase and kill the host tissues, depends on the type of interaction and mostly on the isolates' aggressiveness.

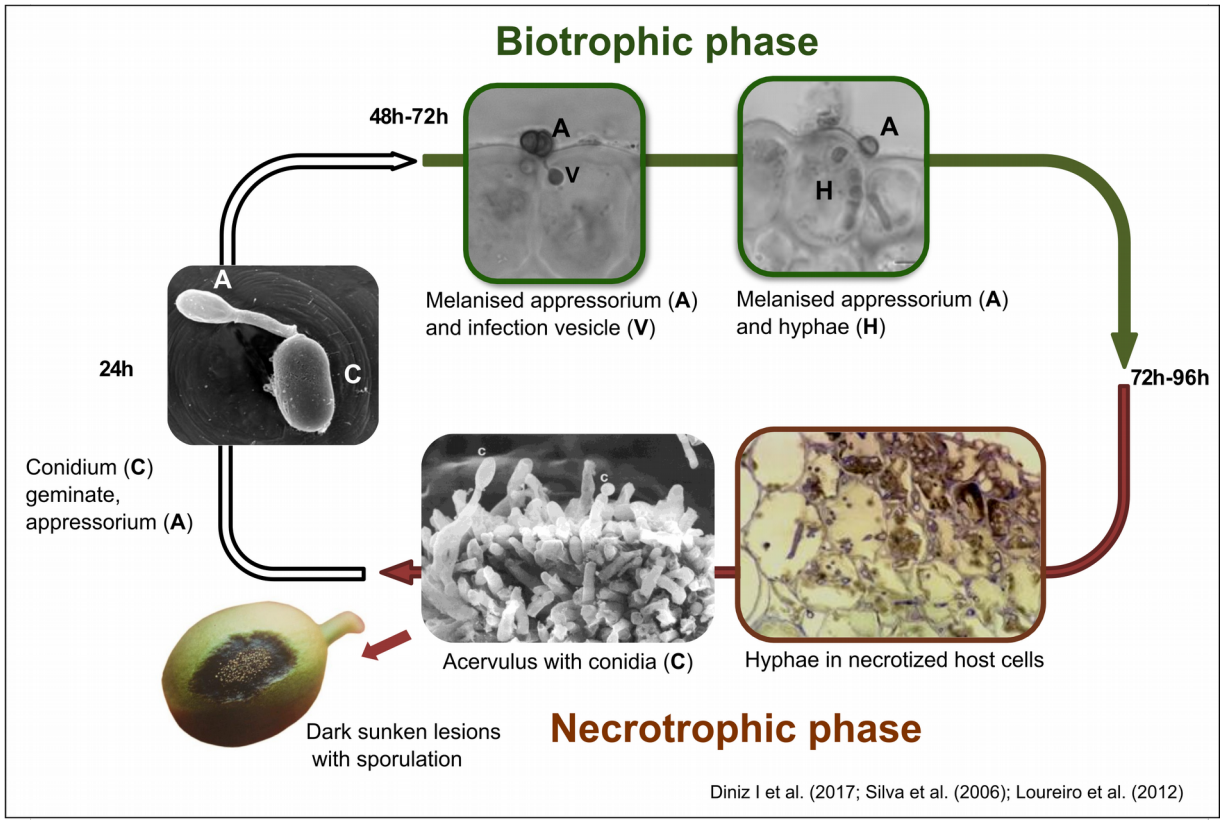


Figure 1.3 - Schematic representation of the infection process of *C. kahawae* in coffee arabica

1.2.1.3 Host-pathogen interaction

The *C. arabica* - *C. kahawae* interaction is still poorly studied, especially on the pathogen perspective, being most of the research focused in plant resistance. Currently, none of the *C. arabica's* varieties is completely resistant to *C. kahawae*, and consequently, the coffee resistance to this pathogen is thought be of a quantitative nature. Previous studies showed that this resistance is inheritable, probably controlled by major genes in different loci (Gichuru *et al.*, 2008; Van Der Vossen and Walyaro, 1980), and results from constitutive and induced mechanisms operating at different stages of pathogenesis (Silva et al., 2006 and references within). For instance, the coffee berry cuticle is probably a physical barrier to the penetrating pathogen (Nutman and Roberts, 1960), while the formation of cork barriers is able to block the progress of

fungal invasion (Masaba and Van Der Vossen, 1982). More recently, cytological and biochemical studies revealed that coffee resistance to *C. kahawae* is characterized by restricted fungal growth associated with several host responses, such as hypersensitive-like cell death (HR), callose deposition around intracellular hyphae, accumulation of phenolic compounds, lignification of host cell walls and increased activity of oxidative enzymes (Diniz *et al.*, 2017; Diniz *et al.*, 2018; Loureiro *et al.*, 2012; Silva *et al.*, 2006). On the other hand, a gene expression study with candidate genes, showed that the coffee immune system enable the perception of the pathogen attack as soon as the infection process starts, being the expression of resistance and susceptibility conditioned by the magnitude and/or timing of defense responses, in which the earlier and strong induction of jasmonic acid biosynthesis, receptor-related genes and pathogenesis-related genes, strongly contribute for plant defense (Diniz *et al.*, 2017; Diniz *et al.*, 2018).

Overall, the previous genetic, pathological and cytological studies, give us a good view of the main biological features of this harmful pathogen. However, a holistic approach is required to comprehensively understand the evolutionary history and population dynamics of this pathogen, the genetic mechanisms that gave rise to its pathogenicity, and which are the mechanisms that regulate its aggressiveness pattern.

1.3 An integrative approach in pathogen research

Recent technological advances in both high–throughput sequencing (HTS) and computational tools allowed the development of a new era in plant pathology (Grünwald *et al.*, 2016). These technologies continue to advance rapidly and the costs have declined to a point of becoming affordable to sequence genomes and transcriptomes of many individuals within a species (Fonseca *et al.*, 2016; Goodwin *et al.*, 2016; Grünwald *et al.*, 2016). Currently, it is globally recognized that these technological advances as well as the development of down-stream genomics analyses, are changing the face of most areas of biology (Stapley *et al.*, 2010), including evolutionary biology and pathogen research. Therefore, it became more feasible than ever to identify the genetic loci responsible for pathogen adaptive evolution in non-model organisms as well as to

bridge the gaps between molecular biology, evolutionary biology and epidemiology (Plissonneau *et al.*, 2017; Stapley *et al.*, 2010).

Additionally, it has been shown that genome scans for signatures of selection and divergence are rarely followed up by in depth molecular and field work, which are crucial to validate the causal effect. In fact, without this effort, the selected loci will remain hypotheses and cannot advance our general understanding on the mechanisms of pathogen adaptation (Fonseca *et al.*, 2016). It is in this context that the current thesis was conceived, trying to address some central questions, including: How did *C. kahawae* emerged? What are the possible migration routes of *C. kahawae*? Is it probable that *C. kahawae* disperses to out-of Africa? Does *C. kahawae* exchange genetic material with other species? What is the reproductive mode of *C. kahawae*? Is it possible to associate molecular markers with aggressiveness? To answer to these questions, our research is focused on studying the genetic basis of the pathogen adaptation, population adaptive divergence and genetic variation related with aggressiveness, which ultimately will allow the identification of candidate genes to perform future functional analyses.

1.3.1 Population genomics and RAD-sequencing

The term “population genomics” appeared, for the first time, in 1990 in the context of large-scale polymorphism analyses in humans (Ellegren, 2014). Currently this term, combines the concepts and technologies of genomics with the biological questions of population genetics, and consequently, it is often viewed as an extension of population genetics (Luikart *et al.*, 2003). Fundamentally, population genomics aims to distinguish locus-specific effects from genome wide effects, such as demographic history, inbreeding and population structure (Thomson *et al.*, 2010). The increasing availability of complete genome sequences, coupled with the ability to genotype thousands of single nucleotide polymorphisms (SNPs), makes it possible to go far beyond traditional population genetics and radically upgrade the questions that can be addressed (Wellenreuther and Hansson, 2016). Therefore, population genomics is not only a matter of scaling up to increase power for making inferences about populations processes, but also offers a means to study the genomic landscape and variance of allelic diversity within and between populations (Ellegren, 2014). For all these reasons,

population genomics represents a new area of research that portends significant conceptual breakthroughs in how we view the genetics, evolution and emergence of plant pathogens (Grünwald *et al.*, 2016).

The utility of SNPs for population genetics was recognized early as a valuable tool to reveal the evolutionary history of populations, adaptation and genome evolution, especially in non model organism (Brumfield *et al.*, 2003; Leaché and Oaks, 2017). Erstwhile, profiling a large number of SNPs was only feasible for model organism with well-developed genomic resources, and even then, it was a tricky task in which an ascertainment bias within populations could occur (Gautier *et al.*, 2013; Wang *et al.*, 2012). However, today these markers are a reliable and a practical choice when compared with the alternative methods that are more time-consuming and expensive (Leache *et al.*, 2015; Rius *et al.*, 2015). Their informative potential and pervasiveness throughout the genome made them ideal markers to investigate speciation events, historical demography and population structure (Ellegren, 2014; Leaché and Oaks, 2017). All these events leave signatures in the diversity and allelic frequency of SNPs which, when provided in sufficient number, could unveil the complex evolutionary scenario of a pathogen (Pavey *et al.*, 2015) For instance, a particularly tricky aspect of fungal pathogen populations is that they are often neither strictly clonal nor sexual, but can have a mixture of both reproduction modes. If the reproduction mode of an organism is not known, a common approach would be to estimate the amount of linkage disequilibrium among SNPs via the calculation of the index of association (Agapow and Burt, 2001). With the availability of large SNP data, it became possible to use a standardized form of the index of association which can accommodate mixed reproductive modes by performing a clone correction step (Kamvar *et al.*, 2014). This methodology revealed to be particularly successful and has been extensively applied on fungal populations to understand their often complex reproductive patterns (Ali *et al.*, 2014; Gladieux *et al.*, 2014; Pierre Gladieux *et al.*, 2015; Goyeau *et al.*, 2007). Additionally, population genomics can be combined with quantitative trait loci (QTL) mapping, genome-wide association studies (GWAS) and functional analyses, to improve our understanding on the genomic architecture of adaptation and speciation, as well as on the nature of the genes involved in these processes (Manel *et al.*, 2016).

Genome-wide association studies attempts to identify regions harboring SNPs that affect some phenotype or outcome of interest. Most existing GWAS analyses are “single-SNP” analyses, which simply test each SNP, one at a time, for association with the phenotype. Strong associations between a SNP and the phenotype are interpreted as indicating that SNP, or a nearby correlated SNP, likely affects phenotype, and the causal effect should be further functionally studied. Alternatively, a GWA analysis can be seen as a variable selection regression problem, with the SNPs as the covariates in the regression. Bayesian variable selection regression (BVSR) for GWAS, provides a very natural approach, in which the phenotype is treated as the regression response and the SNPs become regression covariates. In this scenario, the goal of identifying genomic regions likely to harbor SNPs affecting phenotype is accomplished by examining the genomic locations of SNPs deemed likely to have nonzero regression coefficients (Guan and Stephens, 2011). BVSR also has the advantage of producing easily interpretable measures of confidence, such as posterior probabilities, in which the individual covariates have nonzero regression coefficients. This is particularly important in a scenario where the phenotypic variations observed are explained by a group of SNPs of small effect (Guan and Stephens, 2011). Nevertheless, it has been shown that the power of GWAS analyses depends on the phenotypic size effect of the causal SNPs which leads to a more frequent detection of large-effect loci (Pardo-Diaz *et al.*, 2015).

Genomic comparative analyses, between pathogenic and non-pathogenic fungi, can be used to unveil the genes that are potentially involved in pathogenicity. In a clonal pathogen, the identification of these genomic regions is challenging, as each adaptive allele that arises, will be linked to every other allele in the genome (Grünwald *et al.*, 2016; Plissonneau *et al.*, 2017; Shapiro *et al.*, 2009). In this sense, if the goal is to distinguish adaptive loci from other fixed mutations in the clonal background, the typical genome-scan may not work, and hence it is crucial to look for the excess of functional changes, through the ratio of non-synonymous and synonymous substitutions (dN/dS) (Grandaubert *et al.*, 2017; Hohenlohe *et al.*, 2011; Plissonneau *et al.*, 2017). This ratio, provides a direct estimate of whether codons are under selective pressure, as a dN/dS ratio higher than 1 constitutes a signal of the action of positive selection, while values significantly lower than 1 indicate purifying selection (Plissonneau *et al.*, 2017).

High-throughput techniques based on restriction site-associated DNA sequencing are enabling the cost-effective simultaneous discovery and genotyping of thousands of genetic markers for any species, including non-model organism (Andrews *et al.*, 2016). Specifically, RADseq technology was one of the most important scientific breakthroughs in the past decade by allowing the development of up to thousands of polymorphic genetic markers in a single experiment (Andrews *et al*., 2016). This technology evolved from pioneering studies based on microarrays hybridization techniques, to score presence/absence of marker regions adjacent to restriction sites in the genome, and later was adapted for next-generation sequencing (NGS) (Baird *et al.*, 2008). Generally, the RADseq protocol (**Figure 1.4**) starts by producing DNA libraries for subsequent sequencing using restriction enzymes that cut the genome at specific motifs, and generate short-read sequences adjacent to the cutting sites, named RADseq loci (Lowry *et al*., 2016). These loci can be located in all areas of the genome (that is, both coding and non-coding regions), and individuals within or between closely related species generally share most loci due to the conservation of cutting sites (Andrews *et al*., 2016). With this approach, RADseq is able to combine tight control over the fragments resulting from the digestion with high coverage sequencing across many individuals, which makes it one of the most reproducible restriction digest-based methods (McCormack *et al*., 2013).

Nevertheless, despite all the benefits provided by this technology, it is well-known that a certain source of error and bias are introduced with this approach. Firstly, RADseq is based on digestion by restriction enzymes, and therefore allele dropout and null alleles may arise when a polymorphism occurs at a recognition site. Secondly, stochastic variation in the PCR stage of NGS is also reported to skew the amplification of one allele more than the other (Andrews *et al*., 2016) Thirdly, errors related to the preferential amplification based on GC content may also occur (Davey *et al*., 2011). The last two points can lead to down-stream genotyping errors, and therefore, this kind of error is expected during data analysis. Thus, researchers need to employ new techniques that minimize the error and maximize the retrieval of informative loci (Mastretta-Yanes *et al*., 2014).

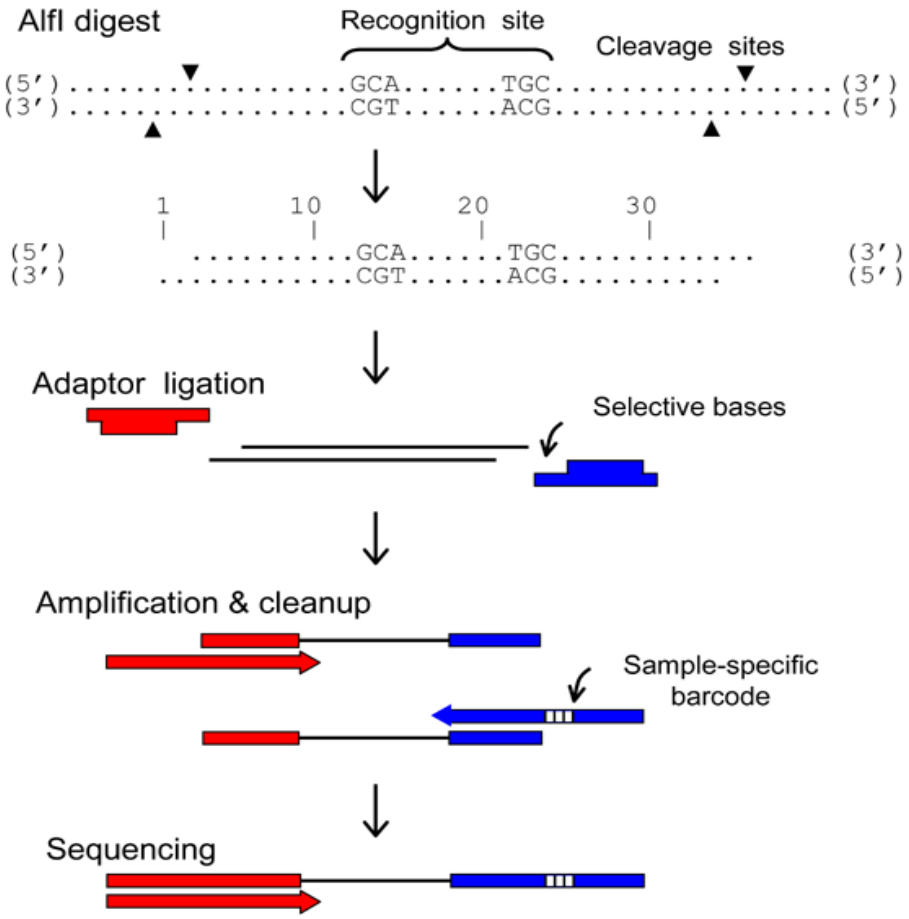


Figure 1.4 - Overview of RAD sequencing library preparation and sequencing following the original protocol. (Retrieved from Wang *et al.* (2012))

Despite that, when implemented appropriately, RADseq, provides an efficient, flexible and cost-effective avenue to unleash the power of NGS for gaining new insights into the evolutionary history and genetic mechanisms underlying the emergence and evolution of populations and species (Andrews *et al.*, 2016).

1.3.2 Pathological studies

Pathological studies allow an accurate phenotypic classification of the most relevant traits in plant pathogens, including the aggressiveness. Economically, this trait is an important factor able to determine the potential ability of a pathogen strain to cause larger or smaller yield losses (Talas *et al.*, 2012). Aggressiveness can be described as a quantitative component of pathogenicity, influenced by genotype and environment, and its role in the adaptation of plant pathogens is still insufficiently investigated (Pariaud *et*

al., 2009). In fact, most of plant-pathogen interaction studies pay attention to the ability of the pathogen to infect the host, and only recently appeared the first studies able to evaluate the quantitative aspects of host-pathogen interactions and their consequences for pathogen evolution. Aggressiveness is traditionally assumed to be polygenically determined, and consequently, it is argued that it is subjected to selection, following a complex pattern in which several trade-offs are made (Boedo *et al*., 2012; Delmas *et al*., 2016). Theoretically, the quantitative adaptation to the host is expected to be slower than the acquisition of additional virulence factors, and the quantitative plant resistance is generally expected to be more durable than qualitative resistance. Therefore, it has been suggested that the quantitative resistance mechanism could be a valuable and durable strategy for crop protection. In this sense, it became crucial to gather empirical knowledge on the nature of aggressiveness, both on its genetic and environmental determinants, to better evaluate the modalities of adaptation of pathogens to their hosts for quantitative traits. Previous studies have already used molecular genetics and genomic approaches to identify the genetic basis underlying pathogenicity and aggressiveness of plant pathogens, including *Blumeria graminis* f. sp. *hordei* (Aguilar *et al.*, 2016), *Fusarium culmorum* (Castiblanco *et al.*, 2018); *Phytophthora infestans* (Cooke *et al.*, 2012); *Fusarium graminearum* (Talas *et al.*, 2016) and *Magnaporthe grisea* (Ebbole, 2007). However, such an approach needs an accurate and comprehensive phenotypic evaluation of pathogen aggressiveness, even more so when genomic association studies are also targeted. In addition, this approach is also crucial to better understand the complex plant-pathogen interactions and can contribute to the deployment of adequate sustainable control strategies.

1.3.3 Follow-up studies using gene expression analysis

The power of population genomics and genome scans to select the loci potentially involved in a given adaptation process is unquestionable. Nevertheless, follow-up studies are crucial to evaluate the causative effect of the selected loci (Pardo-Diaz *et al.*, 2015; Pavey *et al.*, 2010; Tiffin and Ross-Ibarra, 2014). The validation process of a particular selected locus can be time-consuming and technically demanding, but it is a crucial step in finding genes or regulatory elements involved in adaptation (Pardo-Diaz *et al.*, 2015). In this sense, gene expression has grown in popularity to understand

ecological speciation (Pavey *et al.*, 2010) and aggressiveness patterns of pathogens (Aguilar *et al.*, 2016), particularly, when an integrative approach is applied. Currently, mapping SNPs on the reference genome of the studied species or closely related one, can allow the identification of coding gene candidates or loci in physical proximity to coding regions. In that case, if genes are annotated, it can highlight candidate genes functionally consistent with the studied adaptative context (Manel *et al.*, 2016). However, it is important to note that a relation between a candidate gene and a presumptive adaptive phenotype does not constitute the unequivocal detection of a loci responsible for evolutionary change, and therefore, an additional functional analysis (knockouts, knockdowns and transgenics) should be complementarily performed (Pardo-Diaz *et al.*, 2015). Unfortunately, these approaches are quite difficult to implement in non-model organisms and especially in plant pathogens.

Gene expression studies can be conducted at either individual candidate loci or many loci at once, depending on the information and type of resources available (Pardo-Diaz *et al.*, 2015; Pavey *et al.*, 2010). The measure of expression is the abundance of transcribed messenger RNA (mRNA) molecules and this measure is highly correlated with protein abundance (Pavey *et al.*, 2010). The original gene expression technique applied was Northen blotting, but since then, a long way has been covered and today, highly precise and sensitive techniques have been developed (Yan and Liou, 2006). One of these techniques is quantitative real time PCR (qPCR) that has been referred to as one of the most promising approaches to perform gene expression studies, but requires the prior knowledge of candidate genes (Bustin *et al.*, 2010). Basically, the PCR cycle connected to the product exponential growth is tightly associated with the quantity of the initial cDNA template, providing an estimate for the level of mRNA expression in the tissues (Pavey *et al.*, 2010). However, its accuracy is strongly reliant on the use of multiple reference genes for normalization and correction of multiple variation sources. A correct choice of control genes is not a trivial task, and therefore its validation across tissues types and isolates is crucial, as it can affect the accuracy of the calculation of relative expression differences between samples (Pardo-Diaz *et al.*, 2015). Nowadays, high-throughput RNA sequencing technologies (RNA-seq) have the potential to overcome some of these limitations. By using HTS technologies, RNA-seq allows for a direct estimation of relative transcript abundance across the entire genome

without a normalization strategy, but the associated high costs still leave this technology out of reach in many studies (Pardo-Diaz *et al.*, 2015; Pavey *et al.*, 2010). In this work, qPCR technology was used to measure the gene expression of candidate loci, in which a significant effort was made to ensure an accurate data normalization between tissues and different isolates.

In light of the increasing need to control plant pathogens, we consider the close integration of evolutionary genomics with experimental studies to be essential for describing and predicting the emergence, establishment, and adaptation of plant pathogens in agro-ecosystems. The collected information will be important to guide evidence-based sustainable control strategies, aiming to slow down the emergence and spread of pathogens.

1.4 Objectives

The main goal of this thesis was to understand the population genetic structure and dynamics of *C. kahawae,* as well as the adaptive processes responsible for its specific pathogenicity and aggressiveness to *C. arabica*, to contribute for the improvement of the measures currently implemented for coffee protection. Therefore, an integrative approach including population genomics, pathological and gene expression studies was used to specifically address the following objectives:

1. Investigate the demographic and evolutionary history of *C. kahawae* and assess alternative hypothesis on its emergence and dispersal, as well as its evolutionary potential
2. Characterize the aggressiveness profile of a broad range of *C. kahawae* isolates by identifying and quantifying the most relevant parameters to measure the aggressivenesses trait
3. Identify regions across the genome putatively responsible for the pathogenic behavior and aggressiveness of *C. kahawae*, through a genomic comparative analysis and a genome wide association study

4. Establish the first steps for a gene expression follow up study of candidate genes related with *C. kahawae's* pathogenicity and aggressiveness to validate their causal effect

This thesis document comprises six parts. **Chapter 1**, contains a general introduction that contextualizes the overall themes covered in this thesis. **Chapter 2**, addresses the first objective, and contains the results of the demographic and evolutionary history of *C. kahawae*, published in Molecular Plant Pathology. **Chapter 3**, addresses the second objective and contains the results of *C. kahawae* aggressiveness evaluation, already accepted for publication in Plant Pathology. **Chapter 4**, addresses the third objective and describes the detection of genomic regions putatively associated with the pathogenicity and aggressiveness of *C. kahawae*, and subsequent identification of candidate genes. The respective manuscript is in preparation for submission to a peer-review journal. **Chapter 5**, addresses the final objective, and provides the basis to further analyze the candidate genes identified by establishing suitable reference genes for qPCR expression studies in *C. kahawae*. These results were published in PLoS One. Finally, **Chapter 6** contains the conclusion and final remarks of the current thesis.

1.5 References

Agapow, P.-M. and Burt, A. (2001) Indices of multilocus linkage disequilibrium. *Mol. Ecol. Notes* **1**, 101–102.

Aguilar, G.B., Pedersen, C. and Thordal-Christensen, H. (2016) Identification of eight effector candidate genes involved in early aggressiveness of the barley powdery mildew fungus. *Plant Pathol.* **65**, 953–958.

Alemu, K., Adugna, G., Lemessa, F. and Muleta, D. (2017) Current status of coffee berry disease (*Colletotrichum kahawae* Waller & Bridge) in Ethiopia. *Arch. Phytopathol. Plant Prot.* **49**, 421–433.

Ali, S., Gladieux, P., Leconte, M., Gautier, A., Justesen, A.F., Hovmoller, M.S., Enjalbert, J. and Vallavieille-Pope, C. (2014) Origin, migration routes and worldwide population genetic structure of the Wheat Yellow Rust pathogen *Puccinia striiformis* f.sp. *tritici*. *PLoS Pathog.* **10**, e1003903.

Alkimim, E.R., Caixeta, E.T., Sousa, T.V., Pereira, A.A., Carlos, A., Oliveira, B., Zambolim, L. and Sakiyama, N.S. (2017) Marker-assisted selection provides arabica coffee with genes from other *Coffea* species targeting on multiple resistance to rust and coffee berry disease. *Mol. Breed.* ***37****, 6*

Andrews, K.R., Good, J.M., Miller, M.R., Luikart, G. and Hohenlohe, P. a (2016) Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat Rev Genet* **17**, 81–92.

Anthony, F., Combes, M.C., Astorga, C., Bertrand, B., Graziosi, G. and Lashermes, P. (2002) The origin of cultivate*d Coffea arabica* L. varieties revealed by AFLP and SSR markers. *Theor. Appl. Genet.* **104**, 894–900.

Australia Group (2014). Australia Group Common Control List Handbook – Volume II: Biological Weapons-Related Common Control Lists. *Available online at: http://www.australiagroup.net*

Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A. and Johnson, E.A. (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* **3**, 1–7.

Barrés, B., Carlie, J., Seguin, M., Fenouillet, C., Cilas, C. and Ravigné, V. (2012) Understanding the recent colonization history of a plant pathogenic fungus using population genetic tools and Approximate Bayesian Computation. *Heredity (Edinb).* **109**, 269–279.

Batista, D., Silva, D.N., Vieira, A., *et al*. (2017) Legitimacy and implications of reducing *Colletotrichum kahawae* to subspecies in plant pathology. *Front. Plant Sci.* **7**, 2051.

Bazin, É., Mathé-Hubert, H., Facon, B., Carlier, J. and Ravigne, V. (2014) The effect of mating system on invasiveness: some genetic load may be advantageous when invading new environments. *Biol. Invasions* **16**, 875–886.

Bedimo, J.A.M., Bieysse, D., Cilas, C. and Nottéghem, J.L. (2007) Spatio-Temporal dynamics of arabica Coffee Berry Disease caused by *Colletotrichum kahawae* on a plot scale. *Plant Dis.* **91**, 1229–1236.

Beynon, S.M., Coddington, B. and Varzea, V. (1995) Genetic variation in the coffee berry disease pathogen, *Colletotrichum kahawae*. *Physiol. Mol. Plant Pathol.* **46**, 457–470.

Bigger, M. (2006) The dissemination of coffee cultivation throughout the world. *Trop. Agric. Assoc. Newsl* **26**, 15–19.

Boedo, C., Benichou, S., Berruyer, R., *et al*. (2012) Evaluating aggressiveness and host range of *Alternaria dauci* in a controlled environment. *Plant Pathol.* **61**, 63–75.

Bosch, F. van den and Gilligan, C.A. (2008) Models of fungicide resistance dynamics. *Annu. Rev. Phytopathol.* **46**, 123–147.

Bridge, P.D., Waller, J.M., Davies, D. and Buddie, A.G. (2008) Variability of *Colletotrichum kahawae* in relation to other *Colletotrichum* species from tropical perennial crops and the development of diagnostic techniques. *J. Phytopathol.* **156**, 274–280.

Brumfield, R.T., Beerli, P., Nickerson, D.A. and Edwards, S. V. (2003) The utility of single nucleotide polymorphisms in inferences of population history. *Trends Ecol. Evol.* **18**, 249–256.

Burdon, J.J., Thrall, P.H. and Ericson, L. (2013) Genes, communities & invasive species: Understanding the ecological and evolutionary dynamics of host-pathogen interactions. *Curr. Opin. Plant Biol.* **16**, 400–405.

Bustin, S.A, Beaulieu, J.F., Huggett, J., Jaggi, R., Kibenge, F.S.B., Olsvik, P.A, Penning, L.C. and Toegel, S. (2010) MIQE précis: Practical implementation of minimum standard guidelines for fluorescence-based quantitative real-time PCR experiments. *BMC Mol. Biol.* **11**, 74.

Castiblanco, V., Castillo, H. and Miedaner, T. (2018) Candidate Genes for aggressiveness in a natural *Fusarium culmorum* population greatly differ between wheat and rye head blight. *J. Fungi* **4**, 14.

Chung, W., Ishii, H., Nishimura, K., Fukaya, M., Yano, K. and Kajitani, Y. (2006) Fungicide sensitivity and phylogenetic relationship of anthracnose fungi Isolated from various fruit crops in Japan. *Plant Dis.* **90**, 506–512.

Cooke, D.E.L., Cano, L.M., Raffaele, S., *et al*. (2012) Genome analyses of an aggressive and invasive lineage of the Irish Potato Famine pathogen. *PLoS Pathog.* **8**, e1002940.

Crouch, J., O'Connell, R., Gan, P., Buiate, E., Torres, M.F., Beirn, L. and Al., E. (2014) The genomics of *Colletotrichum*, in genomics of plant-associated fungi: Monocot pathogens. *eds Dean R. A., Lichens-Park A., Kole C., Ed. (Berlin Springer-Verlag;)* **69**, 102.

Davey, J.W., Hohenlohe, P.A., Etter, P.D., Boone, J.Q., Catchen, J.M. and Blaxter, M.L. (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* **12**, 499–510.

Delmas, E.L.C., Fabre, F., Jérôme, J., Mazet, I.D., Cervera, S.R., Laurent, D. and François, D. (2016) Adaptation of a plant pathogen to partial host resistance: selection for greater aggressiveness in grapevine downy mildew. *Evol. Appl.* **9**, 709–725.

Derso, E. and Waller, J.M. (2003) Variation among *Colletotrichum* isolates from diseased coffee berries in Ethiopia. *Crop Prot.* **22**, 561–565.

Diniz, I., Azinheira, H., Figueiredo, A., Gichuru, E., Oliveira, H., Guerra-Guimarães, L. and Silva, M.C. (2018) Fungal penetration associated with recognition, signaling and defence-related genes and peroxidase activity during the resistance response of coffee to *Colletotrichum kahawae*. *Physiol. Mol. Plant Pathol.*

Diniz, I., Figueiredo, A., Loureiro, A., *et al*. (2017) A first insight into the involvement of phytohormones pathways in coffee resistance and susceptibility to *Colletotrichum kahawae*. *PLoS One* **12**, 1–20.

Dutech, C., Barre, B., Biology, P., Baillarguet, C.I., Bridier, J., Robin, C., Milgroom, M.G. and Ravignés, V. (2012) The chestnut blight fungus world tour: successive introduction events from diverse origins in an invasive plant fungal pathogen `. *Mol. Ecol.* **doi: 10.11**, 3931–3946.

Dutech, C., Labbé, F., Capdevielle, X. and Lung-Escarmant, B. (2017) Genetic analysis reveals efficient sexual spore dispersal at a fine spatial scale in *Armillaria ostoyae*, the causal agent of root-rot disease in conifers. *Fungal Biol.* **121**, 550–560.

Ebbole, D.J. (2007) *Magnaporthe* as a model for understanding host-pathogen interactions. *Annu. Rev. Phytopathol.* **45**, 437–456.

Elad, Y. and Pertot, I. (2014) Climate Change impacts on plant pathogens and plant diseases. *J. Crop Improv.* **28**, 99–139.

Ellegren, H. (2014) Genome sequencing and population genomics in non-model organisms. *Trends Ecol. Evol.* **29**, 51–63.

Estoup, A. and Guillemaud, T. (2010) Reconstructing routes of invasion using genetic data: why, how and so what? *Mol. Ecol.* **19**, 4113–4130.

Fisher, M.C., Henk, D.A., Briggs, C.J., Brownstein, J.S., Madoff, L.C., McCraw, S.L. and Gurr, S.J. (2012) Emerging fungal threats to animal, plant and ecosystem health. *Nature* **484**, 186–194.

Fonseca, R., Albrechtsen, A., Sibbesen, J.A., Maretty, L., Zepeda-mendoza, M.L., Campos, P.F. and Pereira, R.J. (2016) Next-generation biology: Sequencing and data analysis approaches for non-model organisms. *Mar. Genomics* **30**, 3-13.

Garedew, W., Lemessa, F. and Pinard, F. (2017) Assessment of berry drop due to coffee berry disease and non-CBD factors in Arabica coffee under farmers fields of Southwestern Ethiopia. *Crop Prot.* **98**, 276–282.

Gautier, M., Gharbi, K., Cezard, T., Foucaud, J., Kerdelhué, C., Pudlo, P., Cornuet, J.M. and Estoup, A. (2013) The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Mol. Ecol.* **22**, 3165–3178.

Gibbs, J.N. (1969) Inoculum sources for Coffee Berry Disease. *Ann. Appl. Biol.* **64**, 515–522.

Gichuru, E.K., Agwanda, C.O., Combes, M.C., Mutitu, E.W., Ngugi, E.C.K., Bertrand, B. and Lashermes, P. (2008) Identification of molecular markers linked to a gene conferring resistance to coffee berry disease (*Colletotrichum kahawae*) in *Coffea arabica*. *Plant Pathol.* **57**, 1117–1124.

Giddisa, G. (2016) A Review on the Status of Coffee Berry Disease (*Colletotrichum kahawae*) in Ethiopia. *J. Biol. Agric. Healthc.* **6**, 140–151.

Giraud, T., Gladieux, P. and Gavrilets, S. (2010a) Linking emergence of fungal plant diseases and ecological speciation. *Trends Ecol Evol* **25**, 387–395.

Giraud, T., Gladieux, P. and Gavrilets, S. (2010b) Linking the emergence of fungal plant diseases with ecological speciation. *Trends Ecol. Evol.* **25**, 387–395.

Gladieux, P., Feurtey, A., Hood, M.E., Snirc, A., Clavel, J., Dutech, C., Roy, M. and Giraud, T. (2015) The population biology of fungal invasions. *Mol. Ecol.* **24**, 1969–1986.

Gladieux, P., Ropars, J., Badouin, H., Branca, A., Aguileta, G., Vienne, D.M., Rodriguez de la Vega, R.C., Branco, S. and Giraud, T. (2014) Fungal evolutionary genomics provides insight into the mechanisms of adaptive divergence in eukaryotes. *Mol. Ecol.* **23**, 753–773.

Gladieux, P., Wilson, B.A., Perraudeau, F., *et al*. (2015) Genomic sequencing reveals historical, demographic and selective factors associated with the diversification of the fire-associated fungus *Neurospora discreta*. *Mol. Ecol.* **24**, 5657–5675.

Goodwin, S., McPherson, J.D. and McCombie, W.R. (2016) Coming of age: Ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351.

Goyeau, H., Halkett, F., Zapater, M.F., Carlier, J. and Lannou, C. (2007) Clonality and host selection in the wheat pathogenic fungus *Puccinia triticina*. *Fungal Genet. Biol.* **44**, 474–483.

Grandaubert, J., Dutheil, J.Y. and Stukenbrock, E.H. (2017) The genomic rate of adaptation in the fungal wheat pathogen *Zymoseptoria tritici*. *Doi.Org*, 176727.

Grünwald, N.J., McDonald, B.A.M. and Milgroom, M.G.M.G. (2016) Population Genomics of Fungal and Oomycete Pathogens. *Annu. Rev. Phytopathol.* **54**, 323–346.

Guan, Y. and Stephens, M. (2011) Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Ann. Appl. Stat.* **5**, 1780–1815.

Haag, K.L., Traunecker, E. and Ebert, D. (2013) Single-nucleotide polymorphisms of two closely related microsporidian parasites suggest a clonal population expansion after the last glaciation. *Mol. Ecol.* **22**, 314–326.

Hindorf, H. (1970) *Colletotrichum* spp. isolated from *Coffea arabica* L. in Kenya. *Zeitschrift fur Pflanzenkrankheiten und Pflanzenschutz* **77**, 328–331.

Hindorf, H. and Omondi, C.O. (2011) A review of three major fungal diseases of *Coffea arabica* L. in the rainforests of Ethiopia and progress in breeding for resistance in Kenya. *J. Adv. Res.* **2**, 109–120.

Hohenlohe, P. a., Phillips, P.C. and Cresko, W. a. (2011) Using population genomics to detect selection in natural populations: Key concepts and methodological considerations. *Int. J. Plant Sci.* **171**, 1059–1071.

ICO (2018) World Coffee Production. International Coffee Organization. *http://www.ico.org/prices/po-production.pdf* **[accessed 16 August 2018]**.

Kamvar, Z.N., Tabima, J.F. and Grünwald, N.J. (2014) Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ* **2**, e281.

Kebati, R.K., Nyangeri, J., Omondi, C.O. and Kubochi, J.M. (2016) Effect of artificial shading on severity of Coffee Berry Disease in Kiambu County, Kenya. *Annu. Res. § Rev. Biol.* **9**, 1–11.

Leache, A.D., Chavez, A.S., Jones, L.N., Grummer, J.A., Gottscho, A.D. and Linkem, C.W. (2015) Phylogenomics of phrynosomatid lizards: Conflicting signals from sequence capture versus restriction site associated DNA sequencing. *Genome Biol. Evol.* **7**, 706–719.

Leaché, A.D. and Oaks, J.R. (2017) The utility of Single Nucleotide Polymorphism (SNP) data in phylogenetics. *Annu. Rev. Ecol. Evol. Syst.* **48**, 69–84.

Lécolier, A., Besse, P., Charrier, A., Tchakaloft, T.-N. and Noirot, M. (2009) Unraveling the origin of *Coffea arabica* “Bourbon pointu” from La Réunion: a historical and scientific perspective. *Euphytica* **168**, 1–10.

Loureiro, A., Guerra-Guimarães, L., Lidon, F.C., Bertrand, B., Silva, M.C. and Várzea, V. (2011) Isoenzymatic characterization of *Colletotrichum kahawae* isolates with different levels of aggressiveness. *Trop. Plant Pathol.* **36**, 287–293.

Loureiro, A., Nicole, M.R., Várzea, V., Moncada, P., Bertrand, B. and Silva, M.C. (2012) Coffee resistance to *Colletotrichum kahawae* is associated with lignification, accumulation of phenols and cell death at infection sites. *Physiol. Mol. Plant Pathol.* **77**, 23–32.

Lowry, D.B., Hoban, S., Kelley, J.L., Lotterhos, K.E., Reed, L.K., Antolin, M.F. and Storfer, A. (2016) Breaking RAD: An evaluation of the utility of restriction site associated DNA sequencing for genome scans of adaptation. *Mol. Ecol. Resour* **17**, 142-152.

Luikart, G., England, P.R., Tallmon, D., Jordan, S. and Taberlet, P. (2003) The power and promise of population genomics: from genotyping to genome typing. *Nat. Rev. Genet.* **4**, 981–994.

Luzolo, M., Talhinhas, P., Várzea, V. and Neves-Martins, J. (2010) Characterization of *Colletotrichum Kahawae* isolates causing coffee berry disease in Angola. *J. Phytopathol.* **158**, 310–313.

Manel, S., Perrier, C., Pratlong, M., Abi-Rached, L., Paganini, J., Pontarotti, P. and Aurelle, D. (2016) Genomic resources and their influence on the detection of the signal of positive selection in genome scans. *Mol. Ecol.* **25**, 170–184.

Manga, M., Bieysse, D., Mouen-Bedimo, J., Akalay, I., Bompard, E. and Berry, D. (1997) Observation sur la diversité de la population de *Colletotrichum kahawae* agent de l'anthracnose des baies du cafeier Arabica. Implications pour l'am? eloration getique. *Proc. 17th Int. Conf. Coffee Sci. 1997, Nairobi, Kenya. Paris, Fr. Assoc. Sci. Int. du Cafe*, 604–12.

Masaba, D.M. and Vossen, H.A.H. Van Der (1982) Evidence of cork barrier formation as a resistance mechanism to berry disease (*Colletotrichum coffeanum*) in arabica coffee. *Netherlands J. Plant Pathol.* **88**, 19–32.

Mastretta-Yanes, A., Arrigo, N., Alvarez, N., Jorgensen, T.H., Piñero, D. and Emerson, B.C. (2014) Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Mol. Ecol. Resour.* **15**, 28-41.

McCormack, J.E., Hird, S.M., Zellmer, A.J., Carstens, B.C. and Brumfield, R.T. (2013) Applications of next-generation sequencing to phylogeography and phylogenetics. *Mol. Phylogenet. Evol.* **66**, 526–538.

McDonald, B.A. and Linde, C. (2002) Pathogen population genetics, evolutionary potential, and durable resistance. *Annu. Rev. Phytopathol.* **40**, 349–379.

McDonald, B.A. and Stukenbrock, E.H. (2016) Rapid emergence of pathogens in agro-ecosystems: global threats to agricultural sustainability and food security. *Philos. Trans. R. Soc. B Biol. Sci.* **371**, 20160026.

McCook, S. and Vandermeer, J. (2015). The big rust and the Red Queen: Long-term perspectives on Coffee Rust Research. *Phytopathology* **105**, 1164–1173.

Möller, M. and Stukenbrock, E.H. (2017) Evolution and genome architecture in fungal plant pathogens. *Nat. Rev. Microbiol.* **15**, 756–771.

Motteram, J., Lovegrove, A., Pirie, E., Marsh, J., Devonshire, J., Meene, A. van de, Hammond-Kosack, K. and Rudd, J.J. (2011) Aberrant protein N-glycosylation impacts upon infection-related growth transitions of the haploid plant-pathogenic fungus *Mycosphaerella graminicola*. *Mol. Microbiol.* **81**, 415–433.

Nutman, F. and Roberts, F. (1960) Investigations on a disease of *Coffea arabica* caused by a form of *Colletotrichum coffeanum* Noack. II. some factores affecting germination and infection by the pathogen. *Trans. Br. Mycol. Soc.* **43**, 643–659.

Pardo-Diaz, C., Salazar, C. and Jiggins, C.D. (2015) Towards the identification of the loci of adaptive evolution. *Methods Ecol. Evol.* **6**, 445–464.

Pariaud, B., Ravigné, V., Halkett, F., Goyeau, H., Carlier, J. and Lannou, C. (2009) Aggressiveness and its role in the adaptation of plant pathogens. *Plant Pathol.* **58**, 409–424.

Pautasso, M., Döring, T.F., Garbelotto, M., Pellis, L. and Jeger, M.J. (2012) Impacts of climate change on plant diseases-opinions and trends. *Eur. J. Plant Pathol.* **133**, 295–313.

Pavey, S.A., Collin, H., Nosil, P. and Rogers, S.M. (2010) The role of gene expression in ecological speciation. *Ann. N. Y. Acad. Sci.* **1206**, 110–129.

Pavey, S.A., Gaudin, J., Normandeau, E., Dionne, M., Castonguay, M., Audet, C. and Bernatchez, L. (2015) RAD Sequencing highlights polygenic discrimination of habitat ecotypes in the panmictic American eel. *Curr. Biol.* **25**, 1666–71.

Pinard, F., Omondi, C.O. and Cilas, C. (2012) Detached berries inoculation for characterization of coffee resistance to coffee berry disease. *J. Plant Pathol.* **94**, 517–523.

Pires, A.S., Azinheira, H.G., Cabral, A., *et al*. (2016) Cytogenomic characterization of *Colletotrichum kahawae*, the causal agent of coffee berry disease, reveals diversity in minichromosome profiles and genome size expansion. *Plant Pathol.* **65**, 968–977.

Plissonneau, C., Benevenuto, J., Mohd-Assaad, N., Fouché, S., Hartman, F.E. and Croll, D. (2017) Using population and comparative genomics to understand the genetic basis of Effector-Driven fungal pathogen evolution. *Front. Plant Sci.* **8**, 1–15.

Rius, M., Bourne, S., Hornsby, H.G. and Chapman, M. a (2015) Applications of Next-Generation Sequencing to the study of biological invasions. *Curr. Zool.* **61**, 488–504.

Shapiro, B.J., David, L.A., Friedman, J. and Alm, E.J. (2009) Looking for Darwin's footprints in the microbial world. *Trends Microbiol.* **17**, 196–204.

Silva, C., Várzea, V., Guerra-guimarães, L., Azinheira, H.G., Fernandez, D., Petitot, A., Bertrand, B., Lashermes, P. and Nicole, M. (2006) Coffee resistance to the main diseases: leaf rust and Coffee Berry Disease. *Braz. J. Plant Physiol.* **18**, 119–147.

Silva, D.N., Talhinhas, P., Cai, L., Manuel, L., Gichuru, E.K., Loureiro, A., Várzea, V., Paulo, O.S. and Batista, D. (2012) Host-jump drives rapid and recent ecological speciation of the emergent fungal pathogen *Colletotrichum kahawae*. *Mol. Ecol.* **21**, 2655–2670.

Stapley, J., Reger, J., Feulner, P.G.D., *et al*. (2010) Adaptation genomics: the next generation. *Trends Ecol. Evol.* **25**, 705–712.

Talas, F., Kalih, R., Miedaner, T. and McDonald, B.A. (2016) Genome-Wide association study Identifies novel candidate genes for aggressiveness, deoxynivalenol production, and azole sensitivity in natural field populations of *Fusarium graminearum*. *Mol. Plant. Microbe. Interact.* **29**, 417-30.

Talas, F., Wurschum, T., Reif, J.C., Parzies, H.K. and Miedaner, T. (2012) Association of single nucleotide polymorphic sites in candidate genes with aggressiveness and deoxynivalenol production in *Fusarium graminearum* causing wheat head blight. *BMC Genet* **13**, 14.

Talhinhas, P., Batista, D., Diniz, I., *et al*. (2017) The coffee leaf rust pathogen *Hemileia vastatrix*: one and a half centuries around the tropics. *Mol. Plant Pathol.* **18**, 1039-1051.

Tao, G., Hyde, K.D. and Cai, L. (2013) Species-specific real-time PCR detection of *Colletotrichum kahawae*. *J. Appl. Microbiol.* **114**, 828–35.

Thomson, R.C., Wang, I.J. and Johnson, J.R. (2010) Genome-enabled development of DNA markers for ecology, evolution and conservation. *Mol. Ecol.* **19**, 2184–2195.

Tiffin, P. and Ross-Ibarra, J. (2014) Advances and limits of using population genetics to understand local adaptation. *Trends Ecol. Evol.* **29**, 673–680.

Várzea, V.M.P., Rodrigues, C.J., Silva, M., Pedro, J. and Marques, D. (1999) High virulence of a *Colletotrichum kahawae* isolate from Cameroon as plant pathology compared with other isolates from other regions. *InProceedings 18th Int. Conf. Coffee Sci. 1999, Helsinki, Finland. Paris, Fr. Assoc. Sci. Int. du Cafe*, 131.

Vossen, H.A.M. Van Der and Walyaro, D.J. (1980) Breeding for resistance to coffee berry disease in *Coffea arabica* L. II. Inheritance of the resistance. *Euphytica* **29**, 777–791.

Vossen, H. Van Der (2009) The cup quality of disease-resistant cultivars of arabica coffee (*Coffea arabica*). *Exp. Agric.* **45**, 323.

Vossen, H. van der, Bertrand, B. and Charrier, A. (2015) Next generation variety development for sustainable production of arabica coffee (*Coffea arabica* L.): a review. *Euphytica* **204**, 243–256.

Waller, J.M., Bridge, P.D., Black, R. and Hakiza, G. (1993) Characterization of the coffee berry disease pathogen, *Colletotrichum kahawae* sp. nov. *Mycol. Res.* **97**, 989–994.

Wang, S., Meyer, E., McKay, J.K. and Matz, M. V (2012) 2b-RAD: a simple and flexible method for genome-wide genotyping. *Nat. Methods* **9**, 808–810.

Weir, B.S., Johnston, P.R. and Damm, U. (2012) The *Colletotrichum gloeosporioides* species complex. *Stud. Mycol.* **73**, 115–180.

Weir, W., Capewell, P., Foth, B., *et al*. (2016) Population genomics reveals the origin and asexual evolution of human infective trypanosomes. *Elife* **5**, 1–14.

Wellenreuther, M. and Hansson, B. (2016) Detecting polygenic evolution: problems, pitfalls, and promises. *Trends Genet.* **32**, 155–164.

Yan, H.-Z. and Liou, R.-F. (2006) Selection of internal control genes for real-time quantitative RT-PCR assays in the oomycete plant pathogen P*hytophthora parasitica. Fungal Genet. Biol.* **43**, 430–8.

Zhan, J., Thrall, P.H. and Burdon, J.J. (2014) Achieving sustainable plant disease management through evolutionary principles. *Trends Plant Sci.* **19**, 570–575.

Zhan, J., Thrall, P.H., Papaïx, J., Xie, L. and Burdon, J.J. (2015) Playing on a pathogen's weakness: Using evolution to guide sustainable plant disease control strategies. *Annu. Rev. Phytopathol.* **53**, 2.1-2.25.

Zhang, B.-D., Xue, D.-X., Wang, J., Li, Y.-L., Liu, B.-J. and Liu, J.-X. (2015) Development and preliminary evaluation of a genome-wide single-nucleotide

polymorphisms resource generated by RAD-seq for the small yellow croaker (*Larimichthys polyactis*). *Mol. Ecol. Resour.* **16,** 755-68.

Novel insights on colonization routes and evolutionary potential of *Colletotrichum kahawae*, a severe pathogen of *Coffea arabica*

2

Vieira A.[a,b,c], Silva DN.[a,b,c], Várzea V.[a,c], Paulo OS.[b], Batista D.[a,b,c]

[a]**CIFC/ISA** - UL, Oeiras, Portugal; [b]**CoBiG2/cE3c/FCUL** - UL, Lisboa, Portugal; [c]**LEAF/ISA -** UL, Lisboa, Portugal

2.1 Abstract

Pathogenic fungi are emerging at an increasing rate on a wide range of host plants, leading to tremendous threats to the global economy and food safety. Several plant pathogens have been considered to be invasive species, rendering large-scale population genomic analyses crucial to better understand their demographic history and evolutionary potential. *Colletotrichum kahawae* is a highly aggressive and specialized pathogen, causing coffee berry disease in Arabica coffee in Africa. This pathogen leads to severe production losses and its dissemination out of Africa is greatly feared. To address this issue, a population genomic approach using thousands of single nucleotide polymorphisms (SNPs) spaced throughout the genome was used to unveil its demographic history and evolutionary potential. The current study confirms that *C. kahawae* is a true clonal pathogen, perfectly adapted to green coffee berries, with three completely differentiated populations (Angolan, Cameroonian and East African). Two independent clonal lineages were found within the Angolan population as opposed to the remaining single clonal populations. The most probable colonization scenario suggests that this pathogen emerged in Angola and immediately dispersed to East Africa, where these two populations began to differentiate, followed by the introduction in Cameroon from an Angolan population. However, the differentiation between the two Angolan clonal lineages masks the mechanism for the emergence of the Cameroonian population. Our results suggest that *C. kahawae* is completely differentiated from the ancestral lineage, has a low evolutionary potential and a low dispersion ability, with human transport the most likely scenario for its potential dispersion, which makes the fulfilment of the quarantine measures and management practices implemented crucial.

2.2 Introduction

Pathogenic fungi have been emerging at an increasing rate over the last decade on a wide range of host plants and are responsible for many of the world's most devastating plant diseases, posing a major threat to the global economy and food safety (Figueroa *et al.*, 2016; Jones *et al.*, 2008; Plissonneau *et al.*, 2017; Stukenbrock and Bataillon, 2012). Several fungal plant pathogens are considered to be invasive species, i.e. a set of individuals that have been introduced into a new area, have established themselves, increased in number and spread geographically (Estoup and Guillemaud, 2010). An understanding of the dispersal process involved in the introduction of invasive species is of the utmost interest, not only to design suitable quarantine polices, but also to prevent multiple introductions that can boost the evolutionary potential of the pathogen (Barrés *et al.*, 2012). Most of our knowledge on introduction routes is derived from historical and observational data, which are often sparse, incomplete and sometimes misleading. In the last decade, population genetics has proved to be a useful approach for the reconstruction of routes of introduction, highlighting the true story and pointing out the origin of the founder population (Estoup and Guillemaud, 2010). The introduction of fungal plant pathogens generally occurs by two main mechanisms: the transport of infected plant material (human-mediated) and by natural dispersal at different geographical scales (Barrés *et al.*, 2012). However, the true evolutionary potential of a pathogen can be related to several factors, namely gene flow, effective population size, mutation rate, dispersion mode and, more importantly, reproductive mode (McDonald and Linde, 2002; Plissonneau *et al.*, 2017). In an invasion scenario, it has been shown that a mixed reproductive mode can combine the advantages of both sexual and asexual reproduction, leading to important consequences for fungal demographics and evolution in terms of genetic load reduction, fixation of beneficial mutations and adaptation to fluctuating environments (Bazin *et al.*, 2014; Dutech *et al.*, 2012; Dutech *et al.*, 2017; Weir *et al.*, 2016). Therefore, it is crucial to understand the mode of reproduction of a pathogen in order to infer its evolutionary potential and ability to infect new hosts and areas, which sometimes proves to be a challenging task, mainly because the sexual reproduction of the pathogen may be cryptic rare and difficult to observe in the field, leading to misclassifications (Dutech *et al.*, 2010; Haag *et al.*,

2013). Recently, as a result of high-throughput sequencing, the reproduction mode of several pathogens has been reclassified (Plissonneau *et al.*, 2017), proving the ability of population genetics to increase our knowledge of the reproductive system, the complex demographic scenario of the pathogen and the adaptive evolution of plant pathogens in agro-ecosystems (Grünwald *et al.*, 2017; Stukenbrock *et al.*, 2012).

Colletotrichum kahawae Waller and Bridge is a highly aggressive and specialized pathogen, causing coffee berry disease (CBD) in Arabica coffee (*Coffea arabica* L.) particularly at high altitudes (>1400m), throughout the African continent (Batista *et al.*, 2017; Silva *et al.*, 2006). This pathogen belongs to the list of the 10 most important groups of plant-pathogenic fungi in the world, and is listed as a quarantine pathogen in China, Asia and Latin America, and as a biological weapon in Australia (Australia Group, 2014; Batista *et al.*, 2017; Kebati *et al.*, 2016; Tao *et al.*, 2013). Consequently, the pathogen's potential dispersal to other Arabica coffee cultivation regions outside Africa is greatly feared. This disease can quickly destroy 50%–80% of the developing green berries if no control measures are applied (Alemu *et al.*, 2017; Bedimo *et al.*, 2007; Kebati *et al.*, 2016; Loureiro *et al.*, 2012).

CBD was first reported in western Kenya in 1922, where it led to the destruction and abandonment of Arabica coffee plantations in parts of the region, followed by its spread to Angola around 1930 and to Cameroon in 1955–1957 (Giddisa, 2016). Despite the little attention received during the early stages of its documented emergence, African coffee growers soon witnessed a swift dissemination of CBD throughout most of the continent (Nutman and Roberts, 1960; Rodríguez *et al.*, 1992). The free movement of coffee plant material from CBD-infected areas and the poor management practices are the two main factors responsible for the dissemination of the disease throughout all important Arabica-growing areas in Africa (Giddisa, 2016; Kebati *et al.*, 2016). Recently, a population genetics study has shown that *C. kahawae* emerged in Angola by a host jump from a closely related generalist group of fungi that is unable to infect green coffee berries (Silva *et al.*, 2012). This generalist group of fungi has been described as *C. kahawae* subspecies (*C. kahawae* ssp. *ciggaro*) based only on the genealogical concordance of the recognition criteria of phylogenetic species (Taylor *et al.*, 2000), as a result of its remarkable genetic similarity to *C. kahawae* (Weir *et al.*, 2012). However,

Batista *et al.*, (2017) have argued that this taxonomic ranking has led to a wave of confusion with practical consequences and biosecurity implications, and also that a breadth of data supports the distinction of these groups as separate taxa. According to Batista *et al*. (2017), these groups should instead be accepted as cryptic species for the following reasons: (i) *C. kahawae* is the only species able to infect green coffee berries, a unique ecological niche, which separates it on a functional level, and population genetic and evolutionary data show that these groups clearly represent distinct and reproductively isolated biological entities (Silva *et al.*, 2012); (ii) morphological, cultural and biochemical characteristics (Waller *et al*., 1993), as well as a combination of genes [glutamine synthetase, mating type 1-2-1, and a fragment of DNA lyase Apn2 (Silva *et al.*, 2012; Weir *et al.*, 2012)] and the number of chromosomes (Pires *et al.*, 2016), allow a complete separation of *C. kahawae* from *C. ciggaro.* Furthermore, *C. ciggaro* may even be an unresolved species complex, as an isolate of this group has already been reassigned to a different species (Doyle et al., 2013), reflecting the state of flux of species designation within the genus.

The genetic variability of *C. kahawae* has been assessed with several molecular markers [multi-locus sequences, isoenzymes, rapid amplification of polymorphic DNAs (RAPDs), amplified fragment length polymorphisms (AFLPs)] and low levels of genetic polymorphism were observed (Bridge *et al.*, 2008; Derso and Waller, 2003; Loureiro *et al.*, 2011; Silva *et al.*, 2012). Nevertheless, Silva *et al.*, (2012) were able to demonstrate a clear population structure consisting of three lineages from Angola, Cameroon and East Africa. The demographic pattern observed suggests a dispersion of the pathogen from Angola to Cameroon, and from there to East Africa, with each population isolated in its respective highland area. However, these demographic inferences were made only on the basis of two loci with three single nucleotide polymorphisms (SNPs). Finally, *C. kahawae* has been characterized as a true clonal pathogen with no evidence of sexual reproduction (Silva *et al.*, 2012).

In this study, we applied a population genomics approach, using a broad range of *C. kahawae* isolates, to improve its phylogeny and to comprehensively investigate its demographic history and evolutionary potential, contributing to the implementation of sustainable and durable disease control strategies. As a first goal, we assessed the best

evolutionary hypotheses for the emergence and spread of this pathogen, revealing its origin and the chronology of the colonization routes. For this, we tested two evolutionary hypotheses which had been proposed previously: one based on historical data in which the East African population is considered to be the ancestral population, followed by Angolan and Cameroonian populations (*Hypothesis 1*), and one based on the demographic history proposed by Silva *et al.* (2012) (*Hypothesis 2*), which places the origin of *C. kahawae* in Angola, followed by its sequential spread to Cameroon and East Africa. In addition, two alternative hypotheses were taken into consideration during this work because of their equal probability of occurrence: *Hypothesis 3*, in which the three *C. kahawae* populations emerged at the same time; and, finally, *Hypothesis 4*, in which Angolan and East African populations diverged, virtually at the same time, and, after that, the Cameroonian population emerged from the Angolan population (**Figure 2.1**). As a second goal, we tested for the presence of recombination within *C. kahawae* and, finally, we examined the evolutionary and dispersal potential of this harmful pathogen, especially with regard to its escape from Africa.

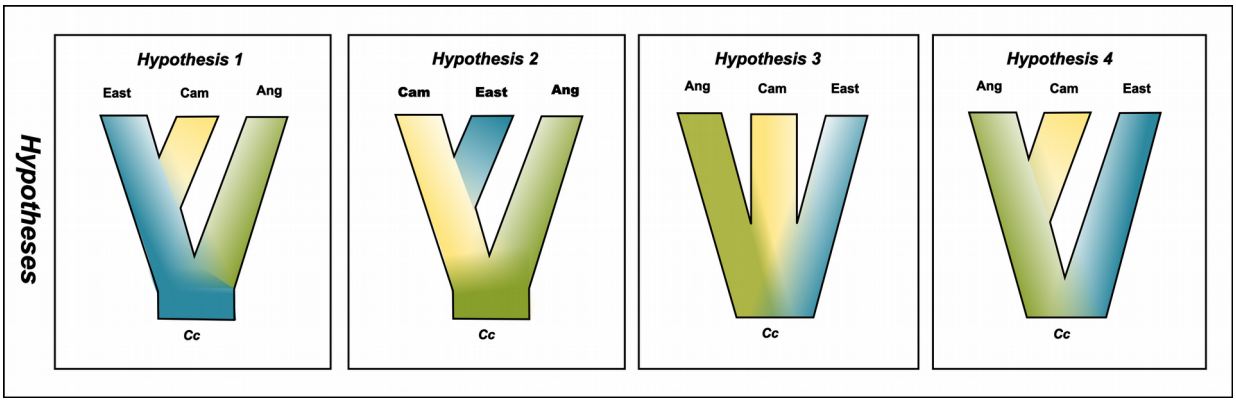


Figure 2.1 - Hypotheses for the colonization scenario of *C. kahawae*. For a more detailed description of the hypotheses initially considered, please see the Introduction section.

2.3 Material and methods

2.3.1 Fungal material, DNA isolation and RAD-seq

In this study, *C. kahawae* ssp. *sensu* Weir *et al.* (2012) is accepted as two cryptic species, as suggested by Batista *et al.* (2017), and described accordingly. Thirty *C. kahawae* (CIFC/ISA/ULisboa collection), collected across the years (1988–2010) from infected green *C. arabica* L. berries in 10 different African countries, were used (**Table A1.1**). These isolates represent the three genetic groups previously described by Silva *et al.* (2012). Moreover, five isolates from the ancestral lineage (*C. ciggaro*), collected from different hosts and countries, were used (**Table A1.1**). Culture and DNA extraction from fungal isolates were performed as described previously by Silva *et al.* (2012), with slight modifications. Briefly, isolates were grown in liquid medium containing 3% malt extract and 0.5% peptone, under a photoperiod of 12 h at 22 °C. DNA was extracted from freeze-dried mycelia with a Sigma Plant/Fungi DNA isolation kit (Sigma-Aldrich, Darmstadt, Germany), according to the manufacturer's instructions. Genomic DNA quality was evaluated by agarose gel and quantified using a Thermo Scientific (Waltham, MA, USA) Nanodrop ND-1000 spectrophotometer.

Three micrograms of high-quality genomic DNA per sample were sent to Floragenex Inc. (Eugene, OR, USA) for RAD library preparation and sequencing. Libraries with sample-specific barcode (eight-nucleotide) sequences were produced from DNA digested with *PstI*. RAD-seq pools were 100-bp single-end sequenced in a lane of an Illumina HiSeq 2000 machine. The sequence data were deposited in the European Nucleotide Archive under submission number PRJEB26929.

2.3.2 RAD-seq quality filtering and SNP calling

Sequence reads were de-multiplexed and quality filtered with the process_radtags program from the package Stacks v1.20 (Catchen *et al.*, 2013). Reads with uncalled bases or distance to barcodes higher than unity were removed. Base calls with a Phred score under 20 were converted to Ns, and reads containing more than four Ns were discarded. Barcodes and Illumina adapters were excluded from each read and the length was truncated to 85 bp (-t 85). Additional filtering and *de novo* assembly within

and between individuals to identify loci were performed using the program PyRAD v3.0.5 (Eaton, 2014), because of its ability to handle indels when clustering sequence reads into othologous loci. Clustering in this study was conducted over a range of parameter values described previously, and the assembly that minimized the number of missing data and maximized the number of phylogenetically informative sites was selected and further processed for the remaining analyses. The sequence variants (SNPs) were then exported into a variant call format (VCF) and the 'stacks' information was exported as a loci file. Handling and exploration of alignment data matrices were performed using TriFusion v1.0.0 software (https://github.com/OdiogoSilva/TriFusion).

2.3.3 Reconstructing of *C. kahawae* phylogeny using RAD data

To assess the phylogenetic relationships amongst the isolates, we used a single concatenated alignment that includes loci with SNPs represented in more than 80% of the isolates and a MAF above 5%. Concatenation and conversion of the alignment matrices to the appropriate formats were performed with TriFusion. RAxML v. 8.2 (Stamatakis, 2014) was used on the CIPRES Portal (Miller *et al*., 2010) to perform a maximum likelihood analysis, employing the general time-reversible (GTR) model of nucleotide substitution with the CAT distributed rate heterogeneity. Non-parametric bootstrapping was implemented in the fast bootstrap algorithm of RAxML with 1000 replicates (MLBS). The GTRCAT general time-reversible model of nucleotide evolution was used, with branch support assessed using 1000 non-parametric bootstrap replicates. Bayesian inference was performed using MrBayes v3.2.6 (Ronquist *et al.*, 2012) with the GTR + Γ model of sequence evolution. The best-fitting models were determined according to the Akaike information criterion (Posada and Buckley, 2004). Posterior probabilities were generated from 1×10^7 generations, sampling at every 1000th iteration, and the analysis was run three times with one cold and three incrementally heated Metropolis-coupled Monte Carlo Markov chains, starting from random trees. The achievement of the stationary phase and mixing were checked for all parameters using Tracer V1.4, and 1 × 106 generations were discarded as burn-in. Trees from different runs were combined using Log combiner and summarized in a majority rule 50% consensus tree. All trees were viewed in FigTree (https://tree.bio.

ed.ac.uk/software/figtree/) and further edited in Inkscape (https://inkscape.org/pt/). It should be noted that, regardless of the dataset under study (dataset generated with different PyRAD parameters), a similar phylogenetic tree was retrieved. Uncertain or controversial relationships were further analyzed with the neighbour-joining split-tree network or the 'neighbor-net' tree implemented on SplitsTree4, with the uncorrected P distance transformation. This method relaxes the assumption that evolution follows a strictly bifurcating path and allows for the identification of reticulated evolution or incomplete lineage sorting amongst the dataset (Huson and Bryant, 2006).

2.3.4 Population structure of *C. kahawae*

We used two individual-centred approaches to describe the genetic structure of populations without a priori geographical knowledge. PCA was run using the SNPRELATE v1.8.0 R package (Zheng *et al.*, 2012) after filtering non-biallelic loci using the snpgdsPCA function. DAPC was used to infer the genetic structure within *C. kahawae* and *C. ciggaro*, as this method takes into account the multi-locus genotype of each individual and forms clusters based on genetic similarity without considering a model of evolution. This analysis was conducted in the R environment with the package ADEGENETv2.0.1 (Jombart, 2008), and the number of clusters was identified on the basis of the Bayesian information criterion (BIC). The genetic differentiation amongst the populations was also assessed by calculating the overall and distribution of SNP FST values for each population pair following VCFTOOLS v0.1.14 (Danecek *et al.*, 2011) and Arlequin v3.5.2 (Excoffier and Lischer, 2010).

2.3.5 Testing *C. kahawae* sexuality

To explore the type of expansion and reproduction system of *C. kahawae* (clonal, partially clonal and/or sexual), we used an R environment Poppr v2.4.0 specifically designed for the genetic analysis of populations with mixed reproduction (Kamvar *et al*., 2015, 2014). The R script used to perform the current analysis, with a detailed description of the command lines, is available at https://github.com/yanavieira/Demographic_history_paper. Briefly, we used the SNP information present in the VCF filtered file to perform the calculation of clone boundaries and collapsed individuals into clonal groups based on the genetic distances. The

remaining analysis, such as reticulations in minimum spanning networks and index of association, were performed comparing both datasets (the clone-corrected and no clone-corrected datasets).

The reticulations in minimum spanning networks were obtained by calculating the minimum spanning tree several times and returning the set of all edges included in the trees (Kamvar *et al.*, 2015). The index of association (IA) is a measure of multi-locus linkage disequilibrium that is most often used to detect clonal reproduction within organisms that have the ability to reproduce via sexual or asexual processes (Brown *et al.*, 1980; Smith *et al.*, 1993), and was standardized by Agapow and Burt (2001) to allow the use of a large amount of loci. During this study, a random sample locus rd calculation was performed in order to obtain a boxplot with the distribution of rd values.

2.3.6 SNP mapping and annotation

As no genome sequence was available for *C. kahawae* or *C. ciggaro*, we used the genome of the most closely related species within the genus *Colletotrichum* [*C. fruticola* (previously misidentified as *C. gloeosporioides* Nara gc5) (Baroncelli *et al.*, 2016), accession_number (GCA_000319635.1) and reference (SAMN02981487)] to perform the locus mapping after de novo assembly. A copy of the assembled scaffolds was obtained from the Ensembl Genome Browser (useast.ensembl.org/index.html). All the loci were then aligned to the reference genomeusing Bowtie 2.2.1.0 (Langmead and Salzberg, 2012) with the '--very-sensitive-localdefault' settings. Alignments were sorted with SAMTools (Li, 2011; Li *et al*., 2009), and the loci that aligned to more than one location were removed from the analysis. The SNP location, annotation and classification ofmtype of mutation were assessed with a personal script developed in python and available at https://github.com/yanavieira/Demographic_history_paper.

2.3.7 Testing the colonization scenarios of *C. kahawae*

Currently, most demographic studies use population genetic models, such as ABC and SFS, to infer the most probable evolutionary scenario of a pathogen (Barrés *et al*., 2012; Dutech *et al*., 2012; Goss *et al.*, 2014; Scarcelli *et al.*, 2013). However, these evolutionary models assume certain conditions, such as the presence of sexual

reproduction, random mating and pure divergence, which would be severely violated in a true clonal pathogen, such as *C. kahawae* . To overcome this problem, a personal script was developed to measure the genetic diversity between all *C. kahawae* populations, and its ancestral lineage was used. These data, linked with the previous analyses, allowed us to infer which is the most probable evolutionary scenario of *C. kahawae*, even without these statistical analyses.

2.4 Results

2.4.1 De novo assembly of restriction site-associated DNA sequencing (RAD-seq) data

In this work, 30 *C. kahawae* isolates, collected from almost all coffee regions in which CBD occurs, and five isolates from the ancestral lineage), were sequenced using Illumina RAD-seq (**Table A1.1**). An average of $3.82x10^6$ reads per sample was generated, yielding $133.78x10^6$ of 85 bp single-end reads after barcode trimming, cleaning and quality checking. The best *de novo* assembly, minimizing the number of missing data and maximizing the number of phylogenetically informative sites, was obtained with parameters consisting of a *minimum depth of coverage* of 10, *maximum number of low quality* of 4, *clustering threshold* of 0.90, *minimal taxon coverage* of 5 and *maximum shared heterozigosity* of 3. The total number of SNPs detected was 173911, but after an initial filtering step that removed SNPs with less than 80% of the taxa represented and a minor allele frequency (MAF) lower than 5%, the *total_dataset* comprised 27099 SNPs across 15007 loci and 35 isolates. In addition, a *ck_dataset* with only *C. kahawae* isolates was also assembled with 3297 SNPs located in 2527 loci across 30 isolates. These two data sets were further used for the remaining phylogenetic and population genetic analyses.

2.4.2 Phylogenetic analysis

Phylogenetic reconstruction of the *total_dataset* with a concatenated set of 27099 variable SNPs using maximum likelihood and bayesian methods showed similar topologies with congruent branch support values (**Figure 2.2**). These analyses revealed a clear differentiation between all the isolates of the two cryptic species (*C. ciggaro* and

C. kahawae), reinforcing the idea that these two pathogens should be completely distinguished taxonomically. Within *C. kahawae* a geographical structure was confirmed with isolates being clustered in three well-supported populations (Angolan, Cameroonian and East African). The Angolan and Cameroonian populations were more closely related to each other than to the East African population. In addition, two well-supported subgroups were observed in Angola, but no other clear substructure was observed. Indeed, bootstrap and posterior probability support values for branches within Cameroonian and East African populations were generally low, which prevents the emergence of any additional substructure.

2.4.3 Population structure and divergence

The *total* and *ck* datasets were used to perform the population genetic analyses. The principal component analyses (PCAs) for the *total_dataset* revealed the same pattern, clearly differentiating the two cryptic species without any geographical structure within *C. kahawae*, with the first two axes being able to explain around 80% of the variation (**Figure 2.3**). By contrast, the *ck_dataset* revealed three clusters within *C. kahawae* (Angolan, Cameroonian and East African), with the Angolan cluster subdivided into two subclusters, which did not present sufficient genetic variability to be considered as an independent group in PCAs. Together, the first two axes were able to explain more than 82% of the variation (**Figure 2.3**). The results of the discriminant analysis of principal components (DAPC) completely corroborated the phylogenetic results, allowing a clear genetic differentiation, not only within *C. kahawae*, but also between the two taxa, with the exception of the Angolan substructure (**Figure A1.1**).

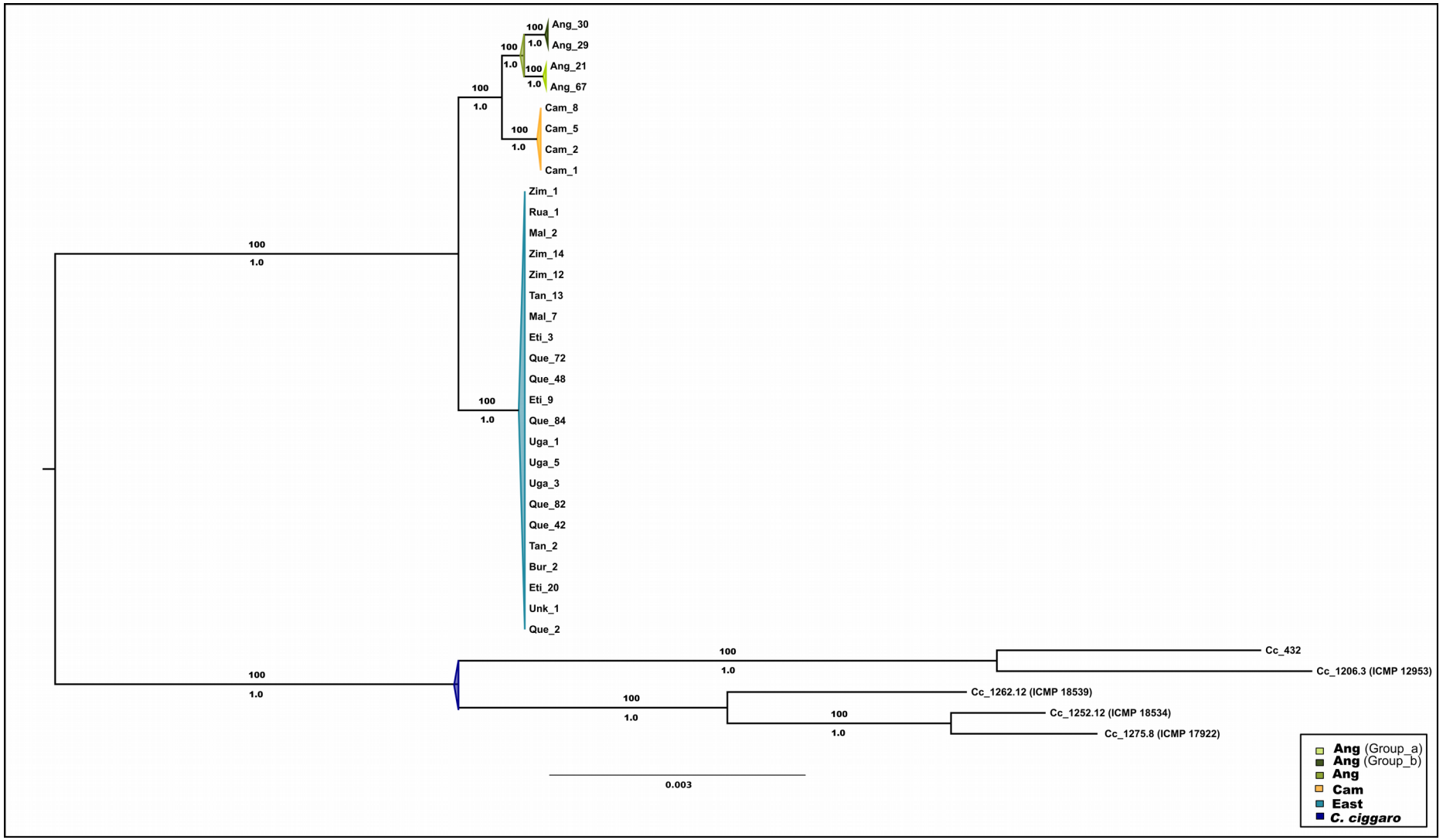


Figure 2.2 - Maximum likelihood phylogenetic tree illustrating the evolutionary relationships amongst the *total_dataset*. Bootstrap and posterior probability values are provided above and below the branches. The three populations within *C. kahawae* [Angolan (Ang), Cameroonian (Cam) and East African (East)], the Angola sub-groups and *C. ciggaro* are shown with different colors.

Estimates of F_{ST} values for each population pair within *C. kahawae* (Angolan, Cameroonian and East African), and between the two cryptic species, further supported a nearly complete genetic differentiation between each group, with weighted F_{ST} values ranging from 0.80 to 0.98 (**Table A1.2**).

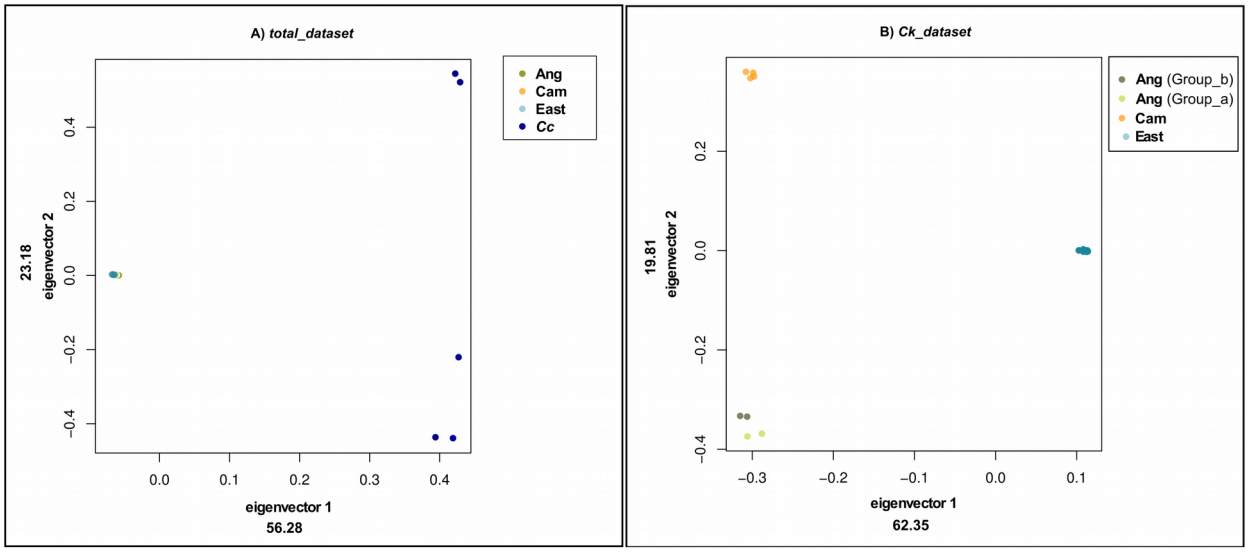


Figure 2.3 - Principal component analysis of genomic diversity for each dataset. **A)** *total_dataset.* **B)** *ck_dataset*. The percentage of variation explained by each principal component is provided in their respective label. Isolates are color coded according to the populations [Angolan (Ang), Cameroonian (Cam) and East African (East)].

2.4.4 Testing potential evolutionary scenarios of *C. kahawae* based on genetic diversity and mapping

A substantial degree of genetic differentiation was observed between *C. kahawae* and its ancestral lineage (*C. ciggaro*), with a total of 9160 SNPs (34%) completely differentiated between them (Fst = 0.89). From this initial dataset, 6641 SNPs had no missing data within *C. kahawae*. As shown in (**Table 2.1**), the Angolan population had a higher number of shared alleles (1045 SNPs) with the ancestral lineage than did the East African (996 SNPs) or Cameroonian (913 SNPs) populations. A similar pattern was observed for the divergent alleles, in which the Cameroonian population (847 SNPs) was the most divergent population when compared with the ancestral lineage, followed by the East African (764 SNPs) and Angolan (715 SNPs) populations. Indeed, no substantial differences in the number of shared/divergent alleles with the ancestral lineage were observed between Angolan and East African populations.

Table 2.1 - Pairwise comparative analyses of the shared and divergent alleles between the three main populations of *C. kahawae* [Angolan (Ang), Cameroonian (Cam) and East African (East)] and between *C. kahawae* and the ancestral lineage

	C. ciggaro	**Ang**	**Cam**	**East African**	
C. ciggaro		715	847	764	
Ang	1045		579	1406	
Cam	913	1181		1535	**Divergent alleles**
East	996	354	225		
	Shared alleles				

The genetic diversity within each population varied greatly. Although the Angolan population had 515 SNPs, the Cameroonian population had 54 SNPs and the East African population had 138 SNPs (**Table 2.2**). A closer examination of the Angolan genetic variation showed that most of the SNPs (432) segregated within two clusters denominated as group_a (Ang_29 and Ang_30) and group_b (Ang_67 and Ang_21). From the remaining 83 SNPs, 29 were variable only in one of the isolates, six SNPs were variable in two random isolates not associated with the established groups and the remaining 46 SNPs contained missing data (**Table 2.2**). The rates of SNPs/individual per population were 20.75, 13.5 and 6.3 for Angolan, Cameroonian and East African populations, respectively, making the East African population the least variable and the Angolan population the most variable.

Table 2.2 - Single nucleotide polymorphism (SNP) variation within each *C. kahawae* population [Angolan (Ang), Cameroonian (Cam) and East African (East)] with and without clone correction

Populations	**SNPs within populations**
Ang	515 (83*)
Cam	138
East	54

* clone corrected SNPs

In agreement with the phylogenetic analysis, Cameroonian and Angolan populations seemed to be more closely related to each other than to the East African population. Briefly, the Angolan/Cameroonian populations shared 1185 SNPs and had 576 divergent alleles, whereas the Angolan/East African populations shared 354 SNPs and

had 1406 divergent SNPs, and the Cameroonian/East African populations shared 225 SNPs and had 1535 divergent SNPs (**Table 2.1**). Included in the divergent SNPs of the Angolan/Cameroonian populations were the 432 SNPs previously associated with the two subgroups within the Angolan population. A more detailed analysis of these SNPs showed that *group_a* had 218 derived alleles, when compared with the ancestral lineage, 53 of which were shared with the Cameroonian population, whereas *group_b* had 207 derived alleles, 30 of which were shared with the Cameroonian population (**Table 2.3**).

Table 2.3 - Single nucleotide polymorphisms (SNPs) segregated within the two Angola clonal lineages

clone_a				**clone_b**			
Ancestral		**Derived**		**Ancestral**		**Derived**	
207		218		217		207	
clone_a = Cam	**clone_a != Cam**	**clone_a = Cam**	**clone_a != Cam**	**clone_b = Cam**	**clone_b != Cam**	**clone_b = Cam**	**clone_b != Cam**
176	30	53	167	165	52	30	179

*The ancestral alleles are the alleles shared with *C. ciggaro* and the clonal lineage and the derived alleles are the SNPs divergent between them. This information was further compared with the Cameroonian population in order to assess the genetic similarity of each clonal lineage with the Cameroonian. != means different

Therefore, it seems that Cameroonian isolates shared information with all Angolan individuals regardless of the group to which they belonged. In order to better understand these results, a neighbour-joining analysis, implemented in SplitsTree, was performed to visualize the possible alternative splits in the dataset. These results were congruent with the previous analyses, but with a possible data conflict split (**Figure 2.4**) in which the Cameroonian population emerged from the Angolan population only after the divergence between the two subgroups.

All of these genetic diversity results strongly suggest that the most probable scenario to explain our results is *Hypothesis 4*; however, the evolutionary mechanism that leads to the emergence of the Cameroonian population is not evident and two new hypotheses can be proposed (**Figure 2.5**).

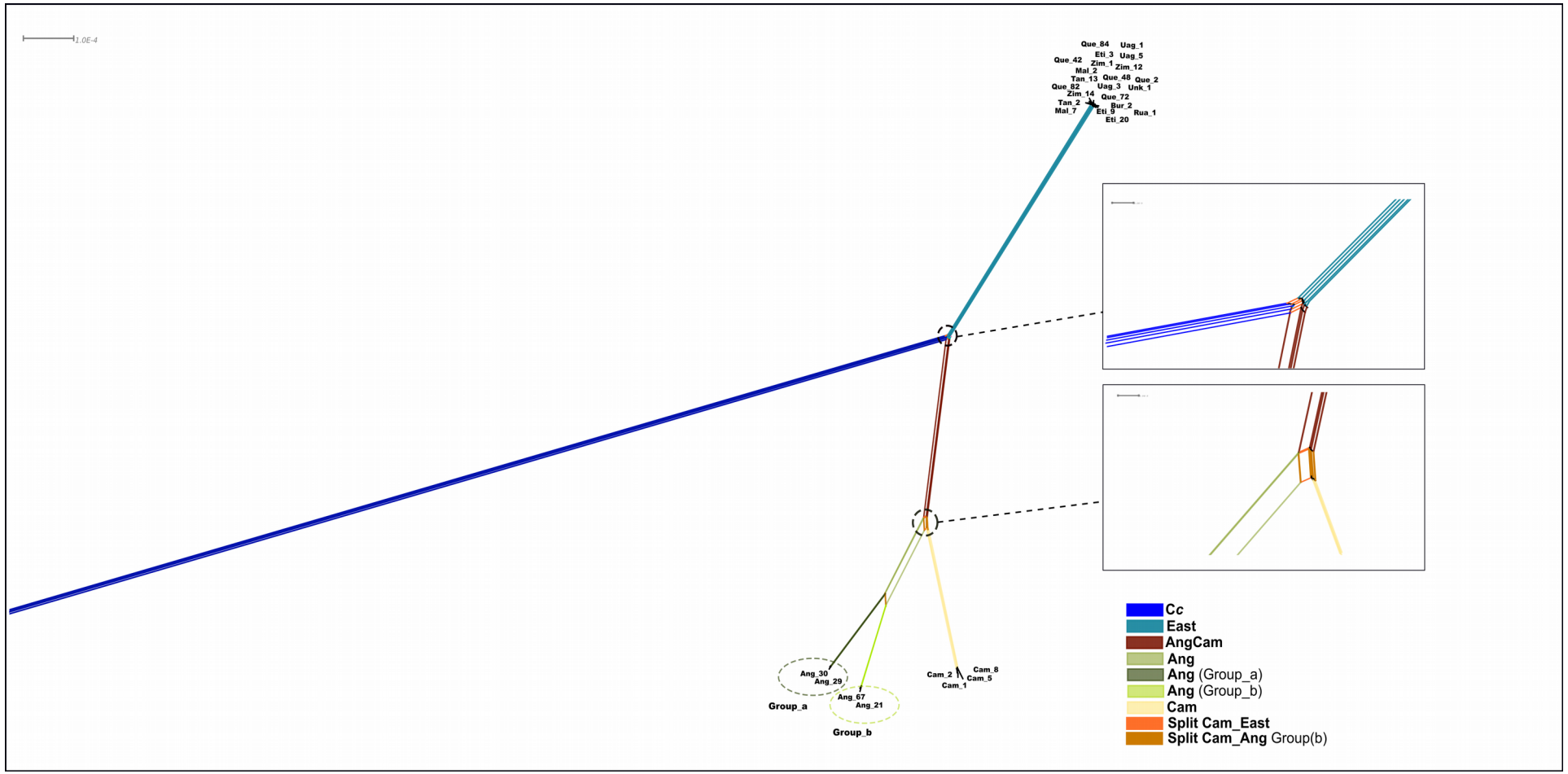


Figure 2.4 - Phylogenetic network inferred using the *total_dataset*. The alternative splits and *C. kahawae* populations are color coded.

Hypothesis 4a suggests that the Angolan population gives rise to the Cameroonian population, and only after that do the two clonal lineages begin to differentiate. *Hypothesis 4b* suggests that the two Angolan clonal lineages begin to differentiate before the emergence of the Cameroonian population, and only after some kind of genetic transfer between them does the Cameroonian population emerge. In order to test the presence of any kind of genetic transmission [hybridization, horizontal gene transfer (HGT) or horizontal whole chromosome transfer (HCT)] without recombination, the segregated SNPs within the two subgroups of the Angolan population were mapped onto the closely related genome available [*C. fruticola* (previously misidentified as *C. gloeosporioides* Nara gc5) (Baroncelli *et al.*, 2016)].

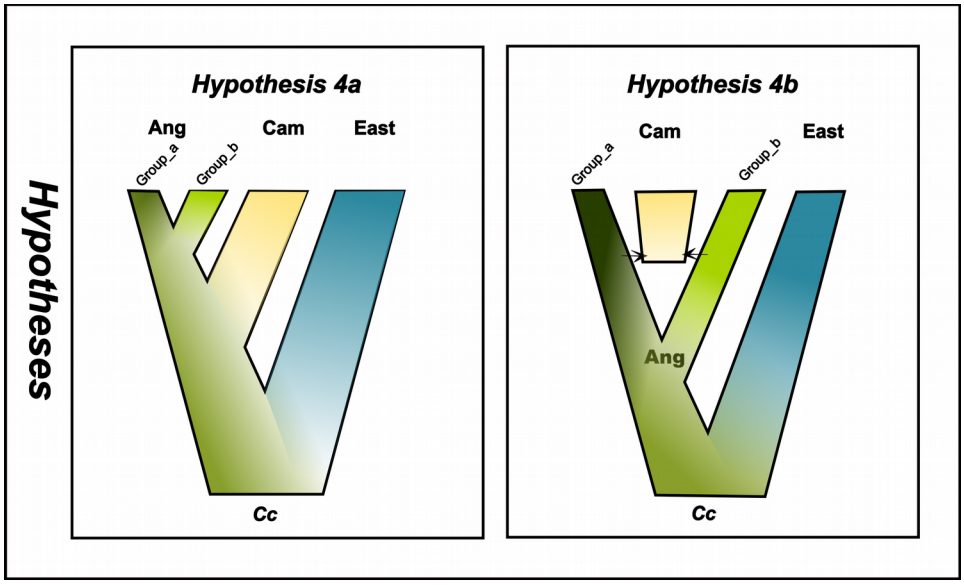


Figure 2.5 - New hypotheses proposed for the colonization scenario of *C. kahawae*. For a more detailed description, please see the Results subsection: 'Testing potential evolutionary scenarios of *C. kahawae* based on genetic diversity and mapping'

From the original 432 segregated SNPs within the two subgroups of the Angolan population, only 84 loci aligned one time, which represents a mapping success of 20%. Within this dataset, 39 loci comprised the SNPs of *group_a* and 45 the SNPs of *group_b*. For *group_a*, only one scaffold (scaffold236) seems to be enriched with five SNPs located in exons (**Table A1.3**). However, when we look only to the SNPs derived from the ancestral lineage and shared between *group_a* and the Cameroonian population (53 SNPs), only 12 were mapped and no scaffold enrichment was detected. Most of these SNPs were located in intergenic regions and only two led to a non-

synonymous mutation. By contrast, for *group_b*, no scaffold enrichment was observed regardless of the type of SNP under study. From the initial dataset of 30 SNPs derived from the ancestral lineage and shared between *group_b* and the Cameroonian population, only 14 were mapped, nine being located in intergenic regions, three responsible for non-synonymous mutations and two for synonymous mutations (**Table A1.3**).

2.4.5 Recombination of *C. kahawae*

We used the Poppr R package to infer the multi-locus genotypes (MLGs) present in *C. kahawae*, and four clonal lineages were detected: two in the Angola, one in Cameron and one in East Africa. The minimum spanning network showed that the two MLGs found in the Angola were the two most closely related clonal lineages, followed by the Cameroon lineage and, finally, the East Africa lineage, in both clone-corrected and uncorrected datasets (**Figure A1.2**).

To explore the type of expansion and reproduction system of *C. kahawae* (clonal, partially clonal and/or sexual) and the presence of recombination, we estimated the standardized form of the index of association (r_d) for both data sets (clone-corrected and uncorrected data sets). In a random mating population under complete linkage equilibrium, r_d, is expected to be zero. As shown in **Figure 2.6**, *C. kahawae* seems to be a true clonal pathogen in which recombination has been absent or extremely rare, regardless of the dataset under-study, with an average value of $r_d = 0.5$. These results were consistently obtained in five independent runs and allowed the rejection of the null hypothesis of linkage equilibrium ($p = 0.001$).

2.5 Discussion

In this work, we used, for the first time, a next-generation sequencing approach to investigate the genetic variability between *C. kahawae* and its ancestral lineage (*C. ciggaro*), and within *C. kahawae*. This approach allowed the generation of genome-wide SNPs, providing the necessary information to perform a comprehensive phylogenetic and population genomic study, and, subsequently, to infer the evolutionary potential, demographic history and current level of adaptive potential of this harmful pathogen.

The collected information will be crucial to understand the disease dynamics, reinforcing the need to respect the implemented quarantine measures and to design future sustainable disease control strategies.

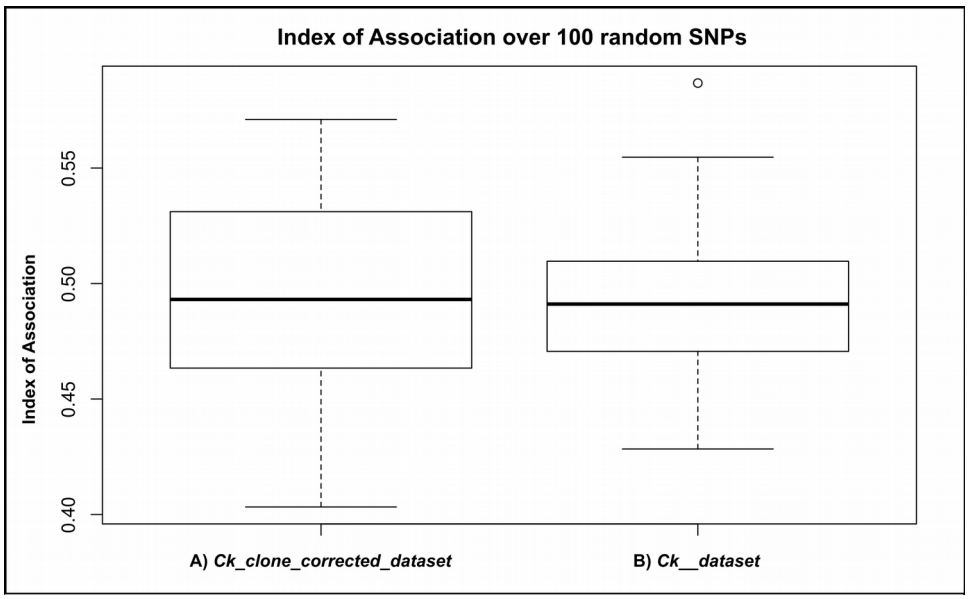


Figure 2.6 - Boxplots with the (r_d) distribution. Two datasets were used: *ck_clone_corrected_dataset* and *ck_dataset*. Each box represents the 100 random samples of 50 variants used to calculate a (r_d) distribution and is centered around the mean, with whiskers extending out to 1.5 times the interquartile range. The median is indicated by the center line.

2.5.1 *C. kahawae* as a distinct invasive species

During this study, an unexpectedly high genetic differentiation between the two cryptic species was observed, with 9160 SNPs completely differentiated between them and a high F_{ST} value (0.89) observed. These results strongly support the idea that these two groups should be considered as different species, especially when the spread of *C. kahawae* out of Africa is so feared and the sequencing of some barcode genes is unable to differentiate them, leading to false reports and widespread panic (Batista *et al.*, 2017). Moreover, the long tree branches within the *C. ciggaro* clade suggest highly divergent groups which may even represent different species; indeed, some members of *C. ciggaro* have been reassigned to the other species (Doyle *et al*., 2013). Despite this, these two cryptic species are genetically close, corroborating the previous hypothesis that *C. kahawae* emerged after a severe bottleneck by a host-jump from this

generalist and widespread group of fungi (Silva *et al*., 2012). Our results, associated with pathological data previously obtained by Silva *et al.*, (2012), corroborate that *C. kahawae* is a successful invader well adapted to its new environment, green coffee berries, in which the evolutionary changes have allowed a rapid and drastic dissemination because of a lack of competition.

2.5.2 Population structure and colonization routes of *C. kahawae*

Phylogenetic and population genetic analyses identified a clear genetic structure within *C. kahawae*, with the East African population being completely differentiated from the Angolan and Cameroonian populations. High genetic similarity was observed between the Angolan and Cameroonian populations, which suggests that these two populations are more closely related to each other than to the East African population. Moreover, evidence of two clonal lineages within the Angolan population was detected, allowing the identification of four independent clonal lineages in *C. kahawae*. In order to infer the chronology of the colonization routes, the number of shared and divergent alleles of each population with the ancestral lineage was measured, revealing that the Angolan population had the highest number of shared alleles with the ancestral lineage, whereas the Cameroonian population had the highest number of divergent alleles. However, no substantial differences were observed between the Angolan and East African populations, suggesting that these two populations probably emerged at a similar time. From these results, two possible scenarios can be conceived: (i) *C. kahawae* could have emerged from an unknown location and reached, at roughly the same time, the Angolan and East African plantations; (ii) the Angolan population could have been established as the emergence starting point of *C. kahawae* and, immediately after, the pathogen was introduced into the East African plantations. Regardless of how the emergence of these two populations occurred, isolation led to a near-complete genetic differentiation between them. Thus, it seems clear that *Hypotheses* 1, 2 and 3 in **Figure 2.1** can be rejected, with *Hypothesis* 4 being the most probable, in which the Angolan and East African populations differentiated, at a similar time, and the Cameroonian population emerged from the Angolan population. These results are not totally concordant with the previous demographic history of *C. kahawae*, in which Angola was

established as the cradle of CBD and, from there, the Cameroonian and, subsequently, East Africa populations emerged (Silva *et al*., 2012). These incongruences are expected as the previous study was based only on two loci and three SNPs. The current results reinforce the importance of genomic studies, comprising a large number of loci, to capture a more accurate demographic history, as results based on a small number of loci can be affected by the lineage sorting history of a particular locus, leading to a biased interpretation of the results.

2.5.3 The emergence of the Cameroonian population

At this point, it seems quite probable that the Cameroonian population emerged from the Angolan population. However, the evolutionary mechanism that led to the appearance of this new population is not evident, especially because of the existence of two clonal lineages in the Angolan population; therefore, two alternative hypotheses are here proposed (*Hypothesis 4a* and *4b*, **Figure 2.5**). The strongest evidence to support *Hypothesis 4a* is the clonality of *C. kahawae*, as no evidence of sexual reproduction or even alternative genetic transmission events was found in our study or referenced in the literature. However, in this scenario, it is very difficult to explain, just by chance, the convergent evolution of several (30 + 53) SNPs from the Cameroonian population with the two clonal lineages of the Angolan population, especially when only one SNP has different genetic information between the three clonal lineages (two Angolan + one Cameroonian) and the ancestral lineage. In contrast, *Hypothesis 4b* implicates the occurrence of some kind of genetic exchange, even if only as a one-time event, which has often been reported in fungi, as a prominent source of adaptation that could have mediated the evolutionary potential of the pathogen. For instance, the acquisition of virulence factors by HCT has been reported in *Fusarium oxysporum*, *Alternaria alternata*, *C. gloeosporioides* and *C. lindemuthianum*, and HGT has been reported in *Pyrenophoratritici-repentis* and *Trichoderma reesi* (Gladieux *et al.*, 2014; Jaramillo *et al.*, 2014; Masel *et al.*, 1996; Mehrabi *et al.,* 2011). Specifically, in *C. gloeosporioides*, it has been shown that a supernumerary 2-Mb chromosome can be transferred between two vegetative incompatible biotypes completely indistinguishable in culture (Masel *et al.*, 1996), and therefore a similar occasional evolutionary event could have occurred in *C. kahawae*. Other evidence that may favour *Hypothesis 4b* is the fact that isolates from

both Angolan clonal lineages were sampled in the same field, which would provide the required conditions for an exchange of genetic material to occur. In an attempt to unveil this conundrum, the loci containing the 432 segregated SNPs within the two Angolan clonal lineages were mapped on the Nara gc5 (*C. fruticola*) genome, the closest related genome available. No evidence of HGT or HCT was detected in all the segregated SNPs or in the specific group of SNPs derived from the ancestral lineage and shared between the Angolan population's *group_a/group_b* and the Cameroonian population. However, caution is needed when analysing these results: (i) only 20% of the SNPs were mapped, leading to a substantial loss of information; (ii) this genome is in a draft stage in which only the scaffold information is available, making it impossible to determine the location of the SNPs on the chromosomes. Therefore, the fact that there is no SNP enrichment in a specific scaffold does not mean that there is no chromosome enrichment. Consequently, in the light of the current information, it is very difficult to infer which of these two evolutionary scenarios would represent more accurately the emergence of the Cameroonian population. In this regard, the sequencing and annotation of the *C. kahawae* genome is of the utmost importance, as the evolutionary potential of this pathogen may change drastically if we find evidence that, in certain unfavourable conditions, *C. kahawae* is able to exchange genetic information between two clonal lineages, and therefore adapt more quickly to a new environment.

2.5.4 The evolutionary and dispersal potential of *C. kahawae*

The results collected during this work when testing the presence of recombination, with and without clone-correction, strongly suggest that *C. kahawae* is a truly clonal pathogen. Even in the case that some singular genetic transmission event may have happened in Angola, according to the definition proposed by Tibayrenc and Ayala (2012), clonality does not necessarily mean a total absence of recombination, only that recombination is rare, which provides support to our inference. In a previous study, *C. kahawae's* low genetic variability associated with a clear genetic structure led to the conclusion that this fungus could be a truly clonal pathogen (Silva *et al*., 2012). Moreover, no vegetative compatibility group was found in *C. kahawae* (He *et al.,* 1998; Várzea, VMP *et al*., 2002). Clonal populations have been shown to have a demographic advantage not only because of the ability of a single individual to colonize an empty

habitat, avoiding the risk of not finding a mating partner, but also because of the ability to produce a large number of offspring and quickly disperse them into an environment to which they are already adapted (Bazin *et al.*, 2014; Dutech *et al.*, 2012, 2017; Hansen *et al.*, 2016). However, the absence of genetic recombination may decrease the adaptation potential in a changing environment, because new genotypes may emerge only by mutation (Dutech *et al.,* 2017). Therefore, it has been argued that completely asexual pathogens are prone to accumulate deleterious mutations and should face extinction in a long-term evolutionary scenario (Bazin *et al.*, 2014; Weir *et al.*, 2016). Consequently, true clonal pathogens are rare, and only a few examples, such as *Hamiltosporidium tvaerminnensis* (Haag *et al.*, 2013), *Verticillium dahliae* (Milgroom *et al.*, 2016) and *Fusarium oxysporum* (Plissonneau *et al.*, 2017), have been described in the literature. Most of the successful invasive pathogens have a dual reproductive system with a clonal phase for dispersion and a sexual phase for gene flow, which deeply increase the evolutionary potential (Gladieux *et al.,* 2015a; Robin *et al.,* 2017). Sexual reproduction or parasexual mechanisms leading to recombination in pathogenic organisms could facilitate the spread of adaptive mutations throughout populations, by creating novel genetic combinations on which selection can act, and promoting the long-term survival of a species (Facon *et al.*, 2008; Haag *et al.*, 2013; Weir *et al.*, 2016).

Nevertheless, it is not only the adaptive genetic potential of a pathogen that matters for a comprehensive risk assessment, but also its ability to disperse the spores and survive (Grünwald *et al.*, 2017). Precipitation, humidity and temperature are the most important factors in CBD outbreaks (Silva *et al.*, 2006). The sporulation of the pathogen is higher at the onset of rain events and the spores are found in mucilaginous masses which prevent them from being dispersed by wind (Giddisa, 2016), but, instead, through ‘splash’ dispersal, which is particularly effective for short-range dissemination (Bedimo *et al.*, 2010). Overall, the dispersal capacity is relatively low, with the exception of passive vectors, such as humans, vehicles, birds and insects, which may carry viable spores, or through the movement of diseased coffee material (unshelled coffee, young plants, green coffee berries and other vegetative materials) (Giddisa, 2016; Kebati *et al.*, 2016). Human intervention seems to be the major factor in the potential dispersal of *C. kahawae*, providing an efficient means of transport over large geographical distances. Another crucial factor that can contribute to to the apparent low evolutionary

potential and dispersal ability is related to the severe changes in the inoculum available during the dry and rainfall seasons: in the dry season, the number of viable spores dramatically decreases, leading to a recurrent bottleneck effect in the population. Given that *C. kahawae* seems to be essentially clonal, spore dispersal is spatially limited and the amount of inoculum available is continuously regulated; consequently, its evolutionary potential is low, which strongly suggests that the ability of this pathogen to overcome plant defenses is low; therefore, an effort to maintain and create new breeding programmes in Africa should be a priority to control this disease. Moreover, if the quarantine practices already in place are correctly followed, such as special care when importing plants from Africa, the use of only dry seeds to circulate plant material and phytosanitary measures, the dissemination of *C. kahawae* outside Africa may be prevented.

In conclusion, despite the advanced methodology used in this study, not all the questions regarding the demographic history of *C. kahawae* could be answered. Our results associated with previously collected data reveal a much more comprehensive demographic history and accurate assessment of the evolutionary potential of this harmful pathogen. The most probable colonization scenario suggests that Angolan and East African populations emerged, virtually at the same time, and the Cameroonian population further emerged from the Angolan population. Despite the most probable scenario placing Angola as the cradle of CBD, the hypothesis that *C. kahawae* could have emerged from an unknown location and subsequently disseminated to the Angolan and East African plantations cannot be rejected. Moreover, for the first time, at least two clonal lineages were detected in the Angolan population, and a scenario in which the exchange of genetic material between these two clones occurred and gave rise to the Cameroonian population cannot also be rejected. Altogether, it has been shown that *C. kahawae* is s a devastating pathogen with a low evolutionary potential and a probably low ability to overcome plant resistance genes, mainly because: (i) it seems to be a true clonal pathogen without any evidence of genetic recombination; (ii) there is an effective regulation of the available inoculum during the dry season, which leads to a recurrent bottleneck effect; and (iii) there is a low dispersion ability, with human transport being the most likely scenario for its dispersion, especially out of Africa. Despite this, evidence indicating that, in certain unfavorable conditions, *C.*

kahawae could exchange genetic information between clonal lineages, and therefore adapt more quickly to a new environment, cannot be ignored and should be tested upon sequencing of the *C. kahawae* genome.

2.6 References

Agapow, P.M. and Burt, A. (2001) Indices of multilocus linkage disequilibrium. *Mol. Ecol. Notes* **1**, 101–102.

Alemu, K., Adugna, G., Lemessa, F. and Muleta, D. (2017) Current status of coffee berry disease (*Colletotrichum kahawae* Waller & Bridge) in Ethiopia. *Arch. Phytopathol. Plant Prot.* **49**, 421–433.

Australia Group (2014). Australia Group Common Control List Handbook – Volume II: Biological Weapons-Related Common Control Lists. *Available online at: http://www.australiagroup.net*

Baroncelli, R., Amby, D.B., Zapparata, A., *et al.* (2016) Gene family expansions and contractions are associated with host range in plant pathogens of the genus *Colletotrichum*. *BMC Genomics* **17**, 555.

Barrés, B., Carlie, J., Seguin, M., Fenouillet, C., Cilas, C. and Ravigné, V. (2012) Understanding the recent colonization history of a plant pathogenic fungus using population genetic tools and Approximate Bayesian Computation. *Heredity (Edinb).* **109**, 269–279.

Batista, D., Silva, D.N., Vieira, A., *et al.* (2017) Legitimacy and implications of reducing *Colletotrichum kahawae* to subspecies in plant pathology. *Front. Plant Sci.* **7**, 1–4.

Bazin, É., Mathé-Hubert, H., Facon, B., Carlier, J. and Ravigne, V. (2014) The effect of mating system on invasiveness: some genetic load may be advantageous when invading new environments. *Biol. Invasions* **16**, 875–886.

Bedimo, J.A.M., Bieysse, D., Cilas, C. and Nottéghem, J.L. (2007) Spatio-Temporal dynamics of arabica coffee berry disease Caused by *Colletotrichum kahawae* on a plot scale. *Plant Dis.* **91**, 1229–1236.

Bedimo, J.A.M., Bieysse, D., Nyasse, S., Nottéghem, J.L. and Cilas, C. (2010) Role of rainfall in the development of coffee berry disease in *Coffea arabica* caused by *Colletotrichum kahawae*, in Cameroon. *Plant Pathol.* **59**, 324–329.

Bridge, P.D., Waller, J.M., Davies, D. and Buddie, A.G. (2008) Variability of *Colletotrichum kahawae* in relation to other *Colletotrichum* species from tropical perennial crops and the development of diagnostic techniques. *J. Phytopathol.* **156**, 274–280.

Brown, A. H., Feldman, M.W. and Nevo, E. (1980) Multilocus structure of natural populations of *Hordeum spontaneum. Genetics* **96**, 523–536.

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A. and Cresko, W. A. (2013) Stacks: An analysis tool set for population genomics. *Mol. Ecol.* **22**, 3124–3140.

Danecek, P., Auton, A., Abecasis, G., *et al.* (2011) The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158.

Derso, E. and Waller, J.M. (2003) Variation among *Colletotrichum* isolates from diseased coffee berries in Ethiopia. *Crop Prot.* **22**, 561–565.

Doyle, V.P., Oudemans, P. V., Rehner, S.A. and Litt, A. (2013) Habitat and host indicate lineage identity in *Colletotrichum gloeosporioides* s.l. from wild and agricultural landscapes in North America. *PLoS One* **8**.

Dutech, C., Barrès, B., Bridier, J., Robin, C., Milgroom, M.G., Ravigné, V. (2012) The chestnut blight fungus world tour: successive introduction events from diverse origins in an invasive plant fungal pathogen. *Mol. Ecol.* **doi: 10.11**, 3931–3946.

Dutech, C., Fabreguettes, O., Capdevielle, X. and Robin, C. (2010) Multiple introductions of divergent genetic lineages in an invasive fungal pathogen, *Cryphonectria parasitica*, in France. *Heredity (Edinb).* **105**, 220–228.

Dutech, C., Labbé, F., Capdevielle, X. and Lung-Escarmant, B. (2017) Genetic analysis reveals efficient sexual spore dispersal at a fine spatial scale in *Armillaria ostoyae*, the causal agent of root-rot disease in conifers. *Fungal Biol.* **121**, 550–560.

Eaton, D.A. (2014) PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics* **30**, 1844–1849.

Estoup, A. and Guillemaud, T. (2010) Reconstructing routes of invasion using genetic data: why, how and so what? *Mol. Ecol.* **19**, 4113–4130.

Excoffier, L. and Lischer, H.E. (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* **10**, 564–567.

Facon, B., Pointier, J.P., Jarne, P., Sarda, V. and David, P. (2008) High genetic variance in life-history strategies within invasive populations by way of multiple introductions. *Curr. Biol.* **18**, 363–367.

Figueroa, M., Upadhyaya, N.M., Sperschneider, J., Park, R.F., Szabo, L.J., Steffenson, B., Ellis, J.G. and Dodds, P.N. (2016) Changing the game: Using integrative genomics to probe virulence mechanisms of the Stem Rust Pathogen *Puccinia graminis* f. sp. *tritici*. *Front. Plant Sci.* **7**, 1–10.

Giddisa, G. (2016) A Review on the Status of coffee berry disease (*Colletotrichum kahawae*) in Ethiopia. *J. Biol. Agric. Healthc.* **6**, 140–151.

Gladieux, P., Feurtey, A., Hood, M.E., Snirc, A., Clavel, J., Dutech, C., Roy, M. and Giraud, T. (2015) The population biology of fungal invasions. *Mol. Ecol.* **24**, 1969–1986.

Gladieux, P., Ropars, J., Badouin, H., Branca, A., Aguileta, G., Vienne, D.M., Rodriguez de la Vega, R.C., Branco, S. and Giraud, T. (2014) Fungal evolutionary genomics provides insight into the mechanisms of adaptive divergence in eukaryotes. *Mol. Ecol.* **23**, 753–773.

Gladieux, P., Wilson, B.A., Perraudeau, F., *et al*. (2015) Genomic sequencing reveals historical, demographic and selective factors associated with the diversification of the fire-associated fungus *Neurospora discreta*. *Mol. Ecol.* **24**, 5657–5675.

Goss, E.M., Tabima, J.F., Cooke, D.E.L., Restrepo, S., Fry, W.E., Forbes, G.A., Fieland, V.J., Cardenas, M. and Grunwald, N.J. (2014) The Irish potato famine pathogen *Phytophthora infestans* originated in central Mexico rather than the Andes. *PNAS* **11**, 8791–8796.

Grünwald, N., Everhart, S., Knaus, B. and Kamvar, Z. (2017) Best practices for population genetic analyses. *Phytopathology* **107***,* 1000-1010*.*

Haag, K.L., Traunecker, E. and Ebert, D. (2013) Single-nucleotide polymorphisms of two closely related microsporidian parasites suggest a clonal population expansion after the last glaciation. *Mol. Ecol.* **22**, 314–326.

Hansen, Z.R., Everts, K.L., Fry, W.E., *et al.* (2016) Genetic variation within clonal lineages of *Phytophthora infestans* revealed through genotyping-by-sequencing, and implications for late blight epidemiology. *PLoS One*, doi:10.1371/journal.pone.0165690.

He, C., Rusu, A.G., Poplawski, A.M., Irwin, J.A.G. and Manners, J.M. (1998) Transfer of a supernumerary chromosome between vegetatively incompatible biotypes of the fungus *Colletotrichum gloeosporioides*. *Genetics* **150**, 1459–1466.

Huson, D.H. and Bryant, D. (2006) Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23**, 254–267.

Jaramillo, V.D.A., Sukno, S.A. and Thon, M.R. (2014) Identification of horizontally transferred genes in the genus *Colletotrichum* reveals a steady tempo of bacterial to fungal gene transfer. *BMC Genet.* **16**, 1–16.

Jones, K.E., Patel, N.G., Levy, M.A., Storeygard, A., Balk, D., Gittleman, J.L. and Daszak, P. (2008) Global trends in emerging infectious diseases. *Nature* **451**, 990-U4.

Jombart, T. (2008) adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403–1405.

Kamvar, Z.N., Brooks, J.C. and Gr??nwald, N.J. (2015) Novel R tools for analysis of genome-wide population genetic data with emphasis on clonality. *Front. Genet.* **6**, 1–10.

Kamvar, Z.N., Tabima, J.F. and Grünwald, N.J. (2014) Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ* **2**, e281.

Kebati, R.K., Nyangeri, J., Omondi, C.O. and Kubochi, J.M. (2016) Effect of artificial shading on severity of coffee berry disease in Kiambu County, Kenya. *Annu. Res. § Rev. Biol.* **9**, 1–11.

Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359.

Li, H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079.

Loureiro, A., Guerra-Guimarães, L., Lidon, F.C., Bertrand, B., Silva, M.C. and Várzea, V. (2011) Isoenzymatic characterization of *Colletotrichum kahawae* isolates with different levels of aggressiveness. *Trop. Plant Pathol.* **36**, 287–293.

Loureiro, A., Nicole, M.R., Várzea, V., Moncada, P., Bertrand, B. and Silva, M.C. (2012) Coffee resistance to *Colletotrichum kahawae* is associated with lignification, accumulation of phenols and cell death at infection sites. *Physiol. Mol. Plant Pathol.* **77**, 23–32.

Masel, A.M., He, C., Popiawski, A.M., Irwin, J.A.G. and Manners, J.M. (1996) Molecular evidence for chromosome transfer between biotypes of *Colletotrichum gloeosporioides*. *Mol. Plant-Microbe Interact.* **9**, 339–348.

McDonald, B.A. and Linde, C. (2002) Pathogen population genetics, evolutionary potential, and durable resistance. *Annu. Rev. Phytopathol.* **40**, 349–379.

Mehrabi, R., Bahkali, A.H., Abd-Elsalam, K.A., *et al*. (2011) Horizontal gene and chromosome transfer in plant pathogenic fungi affecting host range. *FEMS Microbiol. Rev.* **35**, 542–554.

Milgroom, M.G., Mar Jiménez-Gasco, M. del, Olivares-García, C. and Jiménez-Díaz, R.M. (2016) Clonal expansion and migration of a highly virulent, defoliating lineage of *Verticillium dahliae*. *Phytopathology* **106**, 1038–1046.

Miller, M.A., Pfeiffer, W. and Schwartz, T. (2010) Creating the CIPRES Science Gateway for inference of large phylogenetic trees. *2010 Gatew. Comput. Environ. Work. GCE 2010*.

Nutman, F.J. and Roberts, F.M. (1960) Investigations on a disease of *Coffea arabica* caused by a form of *Colletotrichum coffeanum* Noack: I. Some factors affecting infection by the pathogen. *Trans. Br. Mycol. Soc.* **43**, 489–505.

Pires, A.S., Azinheira, H.G., Cabral, A., *et al*. (2016) Cytogenomic characterization of *Colletotrichum kahawae*, the causal agent of coffee berry disease, reveals diversity in minichromosome profiles and genome size expansion. *Plant Pathol.* **65**, 968–977.

Plissonneau, C., Benevenuto, J., Mohd-Assaad, N., Fouché, S., Hartman, F.E. and Croll, D. (2017) Using population and comparative genomics to understand the genetic basis of effector-driven fungal pathogen evolution. *Front. Plant Sci.* **8**, 1–15.

Posada, D. and Buckley, T.R. (2004) Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Syst. Biol.* **53**, 793–808.

Robin, C., Andanson, A., Saint-Jean, G., Fabreguettes, O. and Dutech, C. (2017) What was old is new again: thermal adaptation within clonal lineages during range expansion in a fungal pathogen. *Mol. Ecol.* **doi: 10.11**.

Rodríguez, C.J.., Várzea, V.M.. and Medeiros, E.F. (1992) Evidence for the existence of physiological races of *Colletotrichum coffeanum* Noack sensu Hindorf. *Kenya Coffee (Kenia)* **57**, 1417–1420.

Ronquist, F., Teslenko, M., Mark, P. Van Der, *et al*. (2012) Mrbayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539–542.

Scarcelli, N., Couderc, M., Baco, M.N., Egah, J. and Vigouroux, Y. (2013) Clonal diversity and estimation of relative clone age: application to agrobiodiversity of yam (*Dioscorea rotundata*). *BMC Plant Biol.* **13**, 178.

Silva, C., Várzea, V., Guerra-guimarães, L., Azinheira, H.G., Fernandez, D., Petitot, A., Bertrand, B., Lashermes, P. and Nicole, M. (2006) Coffee resistance to the main diseases: leaf rust and coffee berry disease. *Braz. J. Plant Physiol.* **18**, 119–147.

Silva, D.N., Talhinhas, P., Cai, L., Manuel, L., Gichuru, E.K., Loureiro, A., Várzea, V., Paulo, O.S. and Batista, D. (2012) Host-jump drives rapid and recent ecological speciation of the emergent fungal pathogen *Colletotrichum kahawae*. *Mol. Ecol.* **21**, 2655–2670.

Smith, J.M., Smith, N.H., O'Rourke, M. and Spratt, B.G. (1993) How clonal are bacteria? *Proc.Natl.Acad.Sci.U.S.A* **90**, 4384–4388.

Stamatakis, A. (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313.

Stukenbrock, E.H. and Bataillon, T. (2012) A population genomics perspective on the emergence and adaptation of new plant pathogens in Agro-Ecosystems. *PLoS Pathog.* **8**, 1–4.

Stukenbrock, E.H., Christiansen, F.B., Hansen, T.T., Dutheil, J.Y. and Schierup, M.H. (2012) Fusion of two divergent fungal individuals led to the recent emergence of a unique widespread pathogen species. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 10954–9.

Tao, G., Hyde, K.D. and Cai, L. (2013) Species-specific real-time PCR detection of *Colletotrichum kahawae*. *J. Appl. Microbiol.* **114**, 828–35.

Taylor, J.W., Jacobson, D.J., Kroken, S., Kasuga, T., Geiser, D.M., Hibbett, D.S. and Fisher, M.C. (2000) Phylogenetic species recognition and species concepts in fungi. *Fungal Genet.* Biol. **31**, 21–32.

Tibayrenc, M. and Ayala, F.J. (2012) Reproductive clonality of pathogens: A perspective on pathogenic viruses, bacteria, fungi, and parasitic protozoa. *PNAS* **109**.

Várzea, V.M.P, Rodrigues, J.C. and Lewis, B. (2002) Distinguishing characteristics and vegetative compatibility of *Colletotrichum kahawae* in comparison with other related species from coffee. *Plant Pathol.* **51**, 202–207.

Waller, J.M., Bridge, P.D., Black, R. and Hakiza, G. (1993) Characterization of the coffee berry disease pathogen, *Colletotrichum kahawae* sp. nov. *Mycol. Res.* **97**, 989–994.

Weir, B.S., Johnston, P.R. and Damm, U. (2012) The *Colletotrichum gloeosporioides* species complex. *Stud. Mycol.* **73**, 115–180.

Weir, W., Capewell, P., Foth, B., *et al*. (2016) Population genomics reveals the origin and asexual evolution of human infective trypanosomes. *Elife* **5**, 1–14.

Zheng, X., Levine, D., Shen, J., Gogarten, S.M., Laurie, C. and Weir, B.S. (2012) A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328.

Aggressiveness profiling of the coffee pathogen *Colletotrichum kahawae*

3

Vieira A.[a,b,c], Diniz I.[a,c], Loureiro A.[a,c] , Pereira AP.[a], Silva MC.[a,c] , Várzea V.[a,c] , Batista D.[a,b,c]

[a]**CIFC/ISA** - UL, Oeiras, Portugal; [b]**CoBiG2/cE3c/FCUL** - UL, Lisboa, Portugal; [c]**LEAF/ISA -** UL, Lisboa, Portugal

3.1 Abstract

Colletotrichum kahawae is a specialized plant pathogen of Arabica coffee in Africa, able to infect green berries. The economic impact of this pathogen leads to the urgent need of better understanding its pathogenic lifestyle, namely aggressiveness. In this study, several quantitative traits including disease severity, latent period and incubation period were measured to concomitantly assess the aggressiveness of 26 *C. kahawae* isolates. Our results show that the area under disease progression curve is the most informative variable, particularly when joined together with the index disease intensity 10 days after inoculation and latency period, while the incubation period is not a reliable trait to distinguish aggressiveness levels in *C. kahawae*. This study also confirms the suitability of hypocotyls and detached green berries to perform *C. kahawae* aggressiveness assays, revealing that hypocotyls are a more reproducible testing material. Based on the isolates' profile, three aggressiveness classes were established (*High*, *Moderate* and *Low*). A cytological analysis of representative isolates from each class showed that aggressiveness can be related to the development of post-penetration stages, rather than conidia germination and appressoria differentiation. This study provides, for the first time, the best metrics to evaluate *C. kahawae* aggressiveness, characterizing the profile of a broad range of isolates, and defining a set of parameters that can be used to classify new isolates. Furthermore, the collected information will contribute to improve coffee breeding programs, through the selection of tester isolates for pre-screening of resistant coffee materials and offers the opportunity to engage on future genotype-phenotype studies.

3.2 Introduction

Plant pathologists have focused their attention predominantly on fungal pathogenicity, i.e the qualitative response that reflects the ability of a pathogen to cause disease (Purahong *et al.*, 2012). By contrast, relatively few studies have examined the quantitative aspects involved in host–pathogen interactions, such as pathogen aggressiveness, and their possible consequences at the scale of pathogen populations (Pariaud *et al.*, 2009a) and plant resistance. For several plant pathogen species, aggressiveness is a polygenic quantitative trait, influenced not only by the host genotype but also by the environment (e. g. temperature and humidity) (Boedo *et al.*, 2012; Delmas *et al.*, 2016), and consequently, it is considered to be shaped by natural selection resulting in different adaptive patterns according to the environment (Pariaud *et al.*, 2009b, 2012; Boedo *et al.*, 2012; Delmas *et al.*, 2016). A quantitative adaptation to the host is theoretically expected to be slower than the acquisition of new qualitative virulence factors (Pariaud *et al.*, 2009b). Therefore, it is necessary to gather empirical knowledge on the nature of pathogen aggressiveness, as well as its genetic and environmental determinants, and on the ability of pathogens to respond to the selective pressures imposed by host quantitative resistance (Pariaud *et al.*, 2009b; Delmas *et al.*, 2016). In this study, similar to what has been described in other plant pathogen interaction studies (Boedo *et al.*, 2012; Pariaud *et al.*, 2012; Purahong *et al.*, 2012), “aggressiveness” will be considered as the quantitative measurement of the level of disease reached over time. Therefore, more aggressive pathogens will reach a specific disease level faster than the less aggressive ones. This quantitative trait can be measured using several metrics namely, infection efficiency, latent period, spore production rate, infection period, lesion size and disease severity (Pariaud *et al.*, 2009a; Boedo *et al.*, 2012; Pires *et al.*, 2016), and often vary between species. Thus aggressiveness needs to be evaluated in each particular case (Lee *et al.*, 2015). Moreover, it has been referred that susceptible plants allow the expression of larger aggressiveness differences when exposed to different pathogens isolates (Castiblanco *et al.*, 2018), and consequently, an aggressiveness study should always be performed on a susceptible genotypes.

Colletotrichum kahawae Waller & Bridge, the causal agent of Coffee Berry Disease (CBD), is a specialized hemibiotrophic pathogen of coffee, one of the most important export commodity of tropical countries (Silva *et al.*, 2006; Gichuru *et al.*, 2008; Loureiro *et al.*, 2011). This pathogen currently occurs in nearly all African regions where Arabica coffee (*Coffea arabica* L.) is grown, particularly at high altitudes, ravaging the plantations and causing up to 80% yield losses, if no control measures are applied (Silv*a et al.*, 2006; Bedimo *et al.*, 2010; Hindorf & Omondi, 2011). *C. kahawae* infects all crop organs, from flowers to ripe fruits and occasionally leaves, but maximum crop losses occur due to infection of green berries, causing their premature dropping and mummification (Silva *et al.*, 2006). Due to its huge economic impact, *C. kahawae* is ranked as a quarantine pathogen, namely in Australia and China (Jayawardena *et al*. 2016) and considered as a biological weapon (Australia Group, 2014). Consequently, the pathogen´s potential dispersal to other Arabica coffee cultivation regions is greatly feared, particularly to those at high altitude in Latin America and Asia (Batista *et al.*, 2017). In this sense, a significant effort has been made on the development of resistant varieties to CBD in the scope of breeding programs in Ethiopia, Kenya and Tanzania, as well as on the implementation of preventive selection strategies in Latin America, especially in Colombia (Silva *et al.*, 2006; Van Der Vossen, 2009; Pinard *et al.*, 2012; Alkimim *et al.*, 2017).

Previous studies on *C. kahawae,* have been focused on the variability and differentiation of this pathogen considering morphological, cultural and pathogenic characteristics, vegetative compatibility groups (VCG), DNA sequencing and isoenzymatic methodologies (Omondi *et al.*, 2000; Várzea, VMP *et al.*, 2002; Derso & Waller, 2003; Bridge *et al.*, 2008; Luzolo *et al.*, 2010; Loureiro *et al.*, 2011; Silva *et al.*, 2012). The genetic studies revealed a very low variability within the species, reflecting only a differentiation between three divergent, but clonal, populations (Angolan, Cameroonian and East African) (Silva *et al.*, 2012; Vieira *et al.* 2018). No evidences of *C. kahawae* physiological races were found in *C. arabica*, although differences in the interaction of *C. kahawae* isolates with coffee inter-specific hybrids were reported (Rodrigues *et al.*, 1992; Varzea *et al.,* 1993; Manga *et al.*, 1997). Despite this, significant differences among *C. kahawae* isolates' aggressiveness (i.e number of days until the appearance of first symptoms, number of days until complete necrosis of coffee tissues

and index of disease severity) were early detected and have been referred in the literature (Beynon *et al.*, 1995; Manga *et al.*, 1997; Várzea *et al.*, 1999; Loureiro *et al.*, 2011; Pires *et al.*, 2016; Vieira et *al.*, 2016). Beynon *et al.* (1995) and Derso & Waller (2003) performed the first attempts to associate genetic/molecular markers (RFLPs and RAPDs) with the aggressiveness of *C. kahawae*, but no correlation was found. More recently, Loureiro *et al.* (2011) showed that among six isoenzymatic systems, the alkaline phosphatase was able to discriminate the most and least aggressive isolates of *C. kahawae*. Moreover, Pires *et al.* (2016) suggested a positive relationship between the number of mini-chromosomes and the level of isolate's aggressiveness. However, despite all these efforts, a comprehensive characterization of the aggressiveness of *C. kahawae* using a large set of isolates and metrics, has not yet been performed. On the other hand, validation of the best coffee testing materials for aggressiveness assays is needed. This is extremely relevant not only to define a set of metrics, ranges and conditions able to accurately classify the isolates aggressiveness, but also to create the necessary conditions for developing subsequent studies on the genetic mechanisms underlying these trait, such as genome-wide association studies. Moreover, a comprehensive classification of the isolates aggressiveness will contribute to increase the efficiency of CBD breeding programs, making possible to perform pre-screening resistance tests using isolates that represent the variation of aggressiveness in almost all regions where CBD occurs, instead of using only local isolates. Therefore, the main objectives of the present work were to: i) define the best parameters to measure aggressiveness in Coffee - *C. kahawae* interaction; ii) perform a comparative analysis between hypocotyls and detached green berries for the development of aggressiveness assays; iii) characterize the aggressiveness profiles of 26 *C. kahawae* isolates and establish aggressiveness classes able to accommodate all the variation observed.

3.3 Material and Methods

3.3.1 Fungal isolates

Twenty-six *C. kahawae* isolates (CIFC/ISA/ULisboa collection) from ten different African countries and representative of the three genetic groups previously described by Silva *et al.*, (2012), were used in this study **Table 3.1**. To stimulate conidia production,

isolates were previously cultured on coffee leaf extract agar medium for 7 days under a photo-period of 12 h at 22°C. After that, the isolates were subcultured on 90mm polystyrene Petri dishes containing malt extract agar (MEA, Oxoid (40g/l), England) for 7 days under a photo-period of 12 h at 22°C. Inoculum was obtained by dislodging and harvesting the conidia by flooding the plate with 5 ml of sterile distilled water and the suspensions passed through four layers of sterile muslin cloth to remove mycelia. Concentrations of spore suspensions were determined using a haemocytometer (Weber Scientific Interna-tional, UK) and adjusted to a working concentration of about 2×106 conidia ml-1 for both hypocotyls and green berries inoculation (Loureiro *et al.*, 2011).

Table 3.1 - Details on the *Colletotrichum kahawae* isolates used in this study

C. kahawae Isolates	Geographic origin		Collection year	C. kahawae Isolates	Geographic origin		Collection year
	Country	Region			Country	Region	
Ang 29	Angola	Ganda	2005	**Que 48**	Kenya	Taita Taveta	1996
Ang 6	Angola	Chianga	1992	**Que 82**	Kenya	Kital	2010
Ang 67	Angola	Ganda	2005	**Cam 1**	Cameroon	Babadjou	1992
Zim 12	Zimbabwe	NA	1997	**Cam 2**	Cameroon	Santa	1992
Zim 1	Zimbabwe	Hiton	1991	**Cam 8**	Cameroon	Kumbo	1996
Zim 14	Zimbabwe	NA	1997	**Cam 5**	Cameroon	Baha	1996
Uga 5	Uganda	Kapchorwa	2010	**Tan 12**	Tanzania	Ngoro	2006
Uga 1	Uganda	Kapchorw	2010	**Tan 13**	Tanzania	Mbinga	2006
Uga 3	Uganda	Kapchorwa	2010	**Mal 2**	Malawi	NA	1988
Que 84	Kenya	Mgumguri	2010	**Eti 9**	Ethiopia	Sidamo	1993
Que 2	Kenya	NA	1989	**Eti 20**	Ethiopia	NA	1993
Que 42	Kenya	NA	1996	**Bur 2**	Burundi	NA	1992
Que 72	Kenya	Ruiru	2001	**Rua 1**	Rwanda	Gicumbo	1989

a) NA, not available

3.3.2 Agressiveness assays

Coffee hypocotyls and expanding detached green berries of Arabica coffee var. Caturra (CIFC 19/1; susceptible to all selected *C. kahawae* isolates) were used to perform a comparative aggressiveness trial and assess the best testing system. For that, two independent biological experiments were conducted for each isolate with 30 hypocotyls

and 25 green berries, with a time interval of 3 and 9 months, respectively. Within each biological assay all the isolates were tested simultaneously. The seeds were sown in seedbeds in a growth chamber (FITOCLIMA Walk-in 10000 EHHF, Aralab) under controlled conditions (24–26°C, with 12 h photoperiod at 800 µmol.m-2.s-1 light and 75–85% relative humidity) during 7–8 weeks. Plantlets were collected after emergence (prior to cotyledon expansion) to obtain the hypocotyls structure. The coffee green berries were collected during the berry expanding phase (Mulinge, 1971) from var. Caturra (CIFC 19/1) plants kept in the greenhouse, at temperatures between 16°C and 28°C (average minimum and maximum temperatures, respectively). The detached green berries and hypocotyls were inoculated using the technique described by Loureiro *et al.*, (2011) and Figueiredo *et al.*, (2013), respectively. Briefly, the hypocotyls and green berries were placed on a nylon sponge inside trays, and afterwards the hypocotyls were sprayed, while the green berries were inoculated with a 5µl-drop of conidia suspension. After inoculation, the trays were covered with a plastic bag and maintained in a moist chamber at 22°C in the dark for 24h, and then under a 12h photoperiod during the time-course of the assays.

3.3.3 Definition and assessment of aggressiveness quantitative traits

3.3.3.1 Lesions length and disease severity

Disease symptoms in hypocotyls were scored at 3, 6, 8, 10, 13, 15, 17, 20, 22 and 24 days after inoculation (dai), according to a four-level disease severity scale adapted from Graff (1981): **R0** - No symptoms; **R1** - Discrete necrotic lesions less than 2 mm in length; **R2** - Discrete necrotic lesions less than 6 mm in length; **R3** - Lesions surrounding at least half of the hypocotyl; **R4** - Whole hypocotyls covered with black lesions **Figure 3.1a**. CBD symptoms in green coffee berries were scored for the same time-course referred above, using a different and previously established disease severity scale (Loureiro *et al.*, 2011): **R0** - Green berries without symptoms; **R1** - Black lesions in the inoculation spot (1-2 mm); **R2** - Black lesions with approximately 3mm diameter; **R3** - Black lesions with approximately 5mm diameter; **R4** - Black lesions with approximately 7mm diameter; **R5** - Black lesions with approximately 10 mm diameter; **R6** - Black lesions with approximately 12 mm diameter; **R7** - Black lesions with

approximately 15 mm diameter; **R8** - Whole berries covered with black lesions **Figure 3.1b**. Moreover, to directly compare the quantitative traits of disease severity between the two testing material, the green berries' eight-level scale was converted into a four-level reaction scale by merging two consecutive reaction points (levels) into a new one (**Figure 3.1b**, green scale): **R0** - Green berries without symptoms; **R1** - Black lesions in the inoculation spot (1-3 mm); **R2** - Black lesions with approximately 5-7mm diameter; **R3** - Black lesions with approximately 10-12mm diameter; **R4** - Black lesions covering more than 80% of green berries´ surface (15mm - fully covered with black lesions).

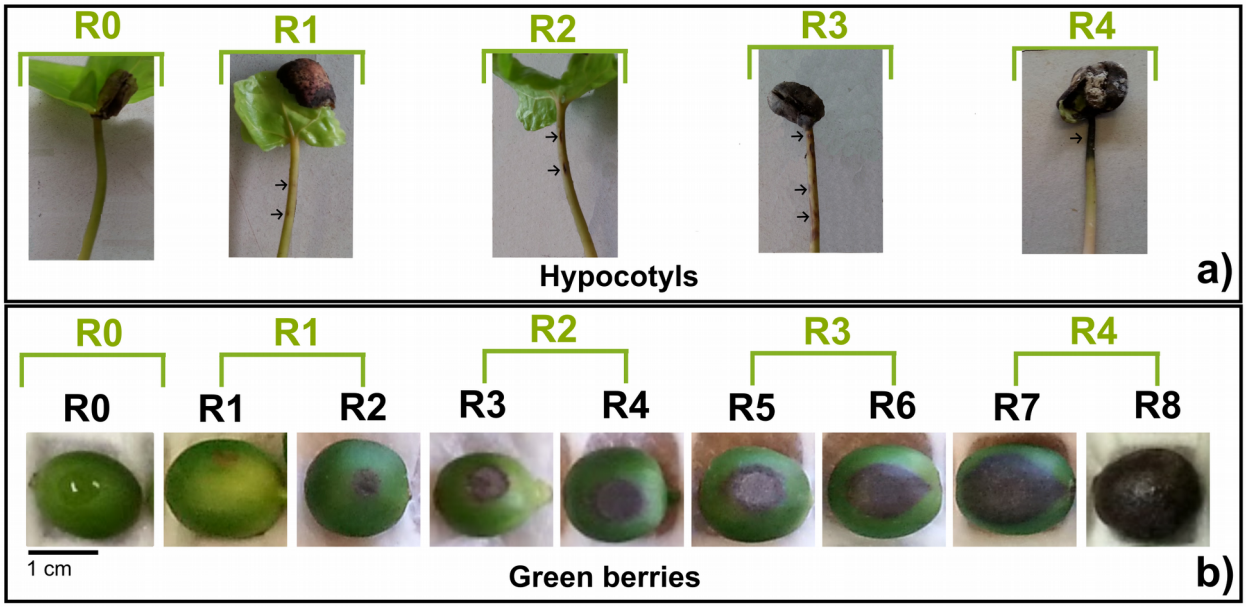


Figure 3.1 - Illustrative scheme of the disease severity scale applied for coffee **a)** hypocotyls and **b)** green berries, comprising the different levels of host reaction (R) to *C. kahawae* throughout the infection time-course. **a)** Four-level scale used to score CBD symptoms in hypocotyls. **b)** Eight-level scale used to score CBD symptoms in green berries (in black); and representation of the four-level merged scale (in green). (Black arrows highlights the symptoms). See scale details on MM :*" Definition and assessment of aggressiveness quantitative traits"*

The data collected from inoculated green berries and hypocotyls were used to calculate the Index of Disease Intensity (*IDI*) (Loureiro *et al.*, 2011) and Disease severity (*DS*) (Pires *et al.*, 2016) during the time course of the assays, using the following formulas:

$$IDI = \sum ¿ \frac{\text{Number of hypocotyls/green berries in each class x Each class ranking}}{\text{Total number of hypocotyls/green berries x Number of classes}}$$

$$DS = \sum ¿ \frac{\text{Number of hypocotyls/green berries in each class x Each class ranking}}{\text{Total number of hypocotyls/green berries}}$$

Symptoms and lesion development along the assessment period were summarized through the computation of the area under the disease progress curves (*AUDPC*) using the *DS* values, and of the standardized area under the disease progression curve (*RAUDPC*) using *IDI* values. Although these two metrics are quite similar, differing only in the normalization by the number of classes included in IDI, both were tested to better assess the most suitable parameters for the characterization of *C. kahawae* aggressiveness profile. Since a positive correlation coefficient value of 1 was observed between them, only the *AUDPC* parameter results are further presented and discussed (**Table A2.1**). Moreover, the IDI at 10dai (*IDI_10dai*) Loureiro *et al.*, (2011) was also recorded in order to assess the disease progression at an early stage of the infection process. This is a standardized variable with values ranging from 0 - 1.

Additionally, other variables were recorded to assess the time taken until reaching the main time points of the infection process, such as: the number of days until more than 50% of the hypocotyls/green berries were completely covered with black lesions (*nr_days_50%*); the number of days to reach severity level 4 (*nr_days_R4*) and the number of days until all hypocotyls/green berries were completely covered with black lesions (*nr_days_100%*).

3.3.3.2 Incubation and latent period

Incubation period was defined as the period of time between inoculation and the appearance of first symptoms (Suffert *et al.*, 2013) , and was recorded as the number of days until the first hypocotyls/green berries showed the first R1 reactions (*nr_days_1stSymptoms*). The latent period was defined as the period of time between inoculation and the appearance of the first conidia (Suffert *et al.*, 2013) and was recorded as the number of days until the first hypocotyls/green berries showed the production of *C. kahawae* conidia (*latent_period*). In our study, the latent period could only be recorded in green berries, as visualization of conidia formation in hypocotyls is not evident and is rapidly superseded by necrotic lesions.

3.3.4 Correlation analyses

The assays reproducibility were assessed using two different approaches: **i)** for each isolate a correlation analysis between the two experimental assay was performed, using

the IDI trait along infection time as a variable. **ii)** for each inoculated plant organ (hypocotyls or green berries), a global correlation analysis between the two experimental assays was performed, using *AUDPC* as a variable. Finally, the comparative analysis between hypocotyls and green berries was performed using the average of *AUPDC* values for each isolate in green berries and in hypocotyls assays, independently. Correlation coefficients were determined using the Pearson correlation at a significant level of 1%.

3.3.5 Group clustering

Group clustering using data from all traits recorded was performed using a heatmap with the hierarchical clustering analysis reduced by Pearson correlation (average linkage clustering) with MeV software (http://mev.tm4.org), not only for the global dataset (green berries and hypocotyls) but also for hypocotyls and green berries specific datasets, which allowed us to group the isolates according to its profile into three main aggressiveness classes (*High*, *Moderate* and *Low*). To perform this statistical analysis, the number of days was normalized in percentage according to the following scale, in which the inoculation day correspond to 0 % and more than 24 dai (>24dai) to 100 %, specifically: 0 dai – 0%; 3 dai - 12%; 6 dai – 24%; 8 dai - 32%; 10 dai – 40%; 13 dai – 52%; 15 dai – 60%; 17dai – 68%; 20 dai –80%; 22 dai – 88%; 24 dai – 96%; >24dai – 100%. The non-parametric Mann-Whitney test *U*, at a significant level of 1%, was used to compare data (*AUDPC*) between the different aggressiveness classes established.

3.3.6 Cytological observations

Conidial germination and appressorial differentiation of the 26 *C. kahawae* isolates was evaluated 1 dai, on lactophenol cotton blue stained nail polish fragments (surface replica) of hypocotyls and green berries (Silva *et al.*, 1999). For each assay, a minimum score of three microscope fields of 100 conidia and/or differentiated appressoria was used. Three isolates (Ang 29, Que 2 and Ang 67), representing the different aggressiveness classes, were chosen to evaluate fungal post-penetration stages. Previous microscope observations showed high similarities on fungal development between green berries and hypocotyls (Silva *et al.*, 2006 and references there in).

Therefore, only cross-sections (20-25 µm) of infected hypocotyls, made with a freezing microtome (Leica CM-1850), were stained and mounted in lactophenol cotton blue (Silva *et al.*, 1999). Hyphal length per infection site was measured with the aid of a micrometric eyepiece. Data was recorded from 75 infection sites per experiment at each time point (1, 2, 3 dai). The observations were made using light microscopy (Leica DM-2500). Finally, the fungal growth was presented as the combined values of three experiments for Ang 29, Que 2 and Ang 67. The Fisher Least Significance Difference (LSD) test was used for statistical analysis.

3.4 Results

3.4.1 Quantitative traits for describing aggressiveness

For the assessment of the aggressiveness of *C. kahawae*, 26 isolates were tested in green berries and hypocotyls of a susceptible *C. arabica* variety. The *AUDPC* values varied significantly among the isolates, ranging from a maximum of 79.42 for hypocotyls and 74.71 for green berries, and a minimum of 40.54 for hypocotyls and 30.44 for green berries (**Table A2.2**). On the other hand, the *IDI_10dai* ranged from 1.0 in the highest aggressive isolates for both inoculated coffee materials, to 0.26 and 0.40 in green berries and hypocotyls, respectively. Additionally, the number of days until more than 50% of the hypocotyls were completely covered with black lesions ranged from 6 to >24dai for the most aggressive and least aggressive isolates respectively, while for green berries it ranged from 8 to >24dai (**Table A2.2**). The same ranges were observed for the number of days required to reach severity level 4 (**Table A2.2**). Finally, all hypocotyls/green berries were completely covered with black lesions between 8 to >24dai for hypocotyls, and 10 to >24dai to green berries (**Table A2.2**). On the other hand, the incubation period ranged from 3 to 10 dai in hypocotyls assays, and from 3 to 8 dai in green berries assays (**Table A2.2**), while the latent period in green berries, ranged from 6 to 13 dai (**Table A2.2**).

The correlation among the eight selected variables is listed in (**Table A2.1)**. In our study, *AUDPC* was the most important variable to evaluate disease dynamics and it was found to be highly positively correlated with *IDI_10dai* in both green berries (r=0.94) and hypocotyls assays (r=0.96), and highly negatively correlated with the remaining time

measurements (*nr_days_50%*; *nr_days_R4*; *nr_days_100%*) (ranging from 0.90-0.96) in both green berries and hypocotyls assays (**Table A2.1**). The latent period showed a moderate negative correlation with *AUDPC* (r= -0.73) and the incubation period did not seem be correlated with *AUDPC* or any of the other quantitative traits (**Table A2.1**), especially in hypocotyls in which the correlation coefficient was always below r= +/- 0.30. The following comparisons between the selected variables (detailed in **Table A2.1**) showed, in general, a high correlation (above r= 0.90) between all aggressiveness quantitative traits studied, with the exception of the latent period that was highly or moderately correlated with the remaining quantitative traits (ranged from r= +/- 0.79-0.67), and the incubation period that was weakly or moderately correlated with the other quantitative traits under study (ranged from r= +/- 0.70 – 0.19).

3.4.2 Validation of experimental reproducibility and comparative analysis between hypocotyls and green berries assays

The assays' reproducibility were assessed at two levels: per isolate, comparing the IDI measured along the infection time; with the *AUDPC* parameter between experimental assays; and per inoculated plant material (green berries and hypocotyls), comparing the global value of AUDPC between the average of experimental assays. Most isolates presented a high germination (average 86%) and appressoria differentiation (average 89%). The correlation coefficient analysis using *IDI* was always high and ranged from r=0.99 to r=0.85 in green berries and r=1.00 to r=0.88 in hypocotyls, which suggests a high inter-assay reproducibility with some variation in the pattern of isolate's reproducibility depending on the testing material (**Table A2.3**). Moreover, the correlation coefficient analysis using the *AUDPC* trait showed a high correlation between hypocotyls assays (r=0.82) and a moderate correlation between detached green berries assays (r=0.59) (**Table A2.4**), suggesting that hypocotyls are a more reproducible material than green berries. Finally, regarding the global comparison between hypocotyls vs green berries assays, a high correlation coefficient value (r=0.77) was observed between the two *C. arabica* organs (**Table A2.4, Figure 3.2)**. Nevertheless, during this analysis some outliers (isolates Que 82, Uga 1, Que 84 and Ang 6) were detected, and as expected when removed from the analyses, a higher correlation

coefficient value was obtained (r=0.81). These results strongly suggest that both testing materials could be used to assess isolate aggressiveness, being hypocotyl a more homogeneous material.

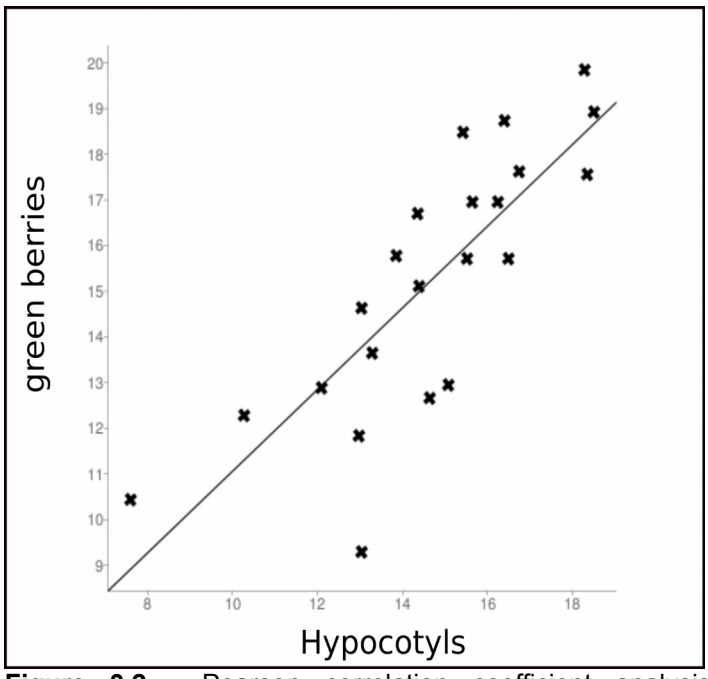


Figure 3.2 - Pearson correlation coefficient analysis between hypocotyls and green berries (r= 0.77; p=<0.00001).

3.4.3 Differentiation of aggressiveness profiles

3.4.3.1 Establishment of aggressiveness classes based on quantitative traits

In order to establish aggressiveness classes, and taking into account the high correlation values observed between the green berries and hypocotyl assays, a heatmap comprising all aggressiveness traits was produced. As shown in **Figure 3.3**, the isolates could be grouped in three different classes of aggressiveness: *High*; *Moderate*; and *Low*. Globally, the classes of aggressiveness established can be described as follows (**Table 3.2**): highly aggressive isolates are able to reach the complete necrosis of coffee tissues in 8 to 10 dai and to reach the severity level 4 between 6 to 9 dai, presenting a *AUDPC* value higher than 74.26 and reaching the *IDI_10dai* maximum value of 1; moderate aggressive isolates are able to reach the complete necrosis of coffee tissues in 15 to 22 dai and to reach the severity level 4

between 10 to 18 dai, while their *AUDPC* value range from 70.25 to 56.68 and their *IDI_10dai* from 0.59 to 0.9; finally, the low aggressive isolates are able to reach the complete necrosis of coffee tissues in more than 22 dai and reach the severity level 4 always after 22 dai, presenting a *AUDPC* value lower than 56.06 and a I*DI_10dai* value always less than 0.57.

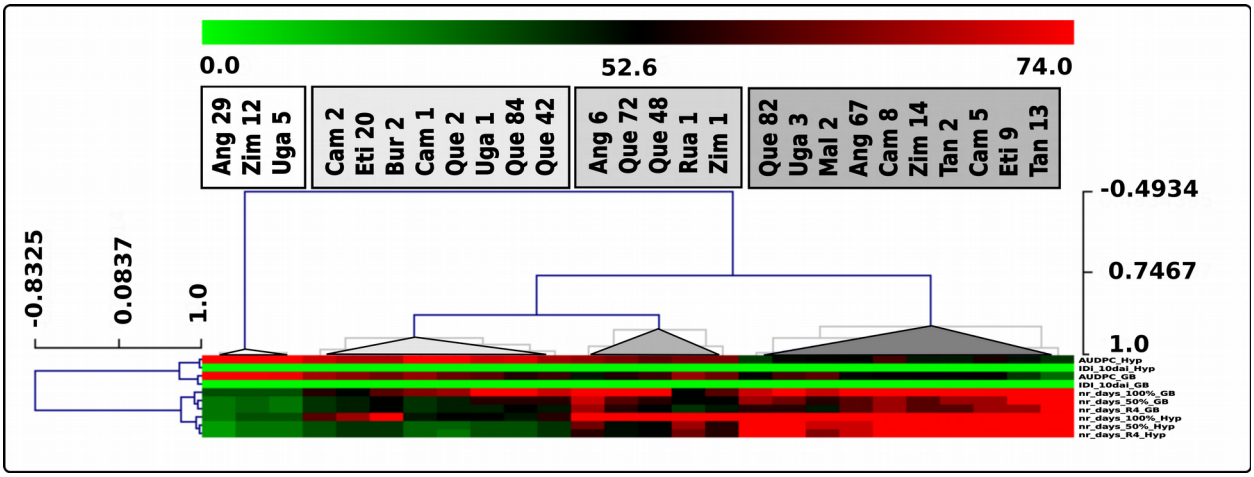


Figure 3.3 - *C. kahawae* isolate group clustering from a heatmap analysis, using the data from all quantitative traits recorded in green berries and hypocotyls. Isolate groups are presented in color coded boxes corresponding to different aggressiveness classes (high - white; high_moderate - light grey; low_moderate - grey; low - dark gray).

Moreover, the moderate class can be sub-divided into two levels: *High_Moderate* and *Low_Moderate*. Specifically, the *High_Moderate* aggressive isolates are able to reach the complete necrosis of coffee tissues in 15 to 18 dai and to reach the severity level 4 between 10 to 13 dai, while their *AUDPC* value range from 70.25 to 62.09 and their *IDI_10dai* from 0.66 to 0.9; *Low_Moderate* aggressive isolates are able to reach the complete necrosis of coffee tissues in 17 to 22 dai and to reach the severity level 4 between 15 to 18 dai, while their *AUDPC* value range from 64.40 to 56.60 and their *IDI_10dai* from 0.59 to 0.75 (**Table 3.2**).

A Mann-Whitney *U* test with *AUDPC* values showed significant differences between the three main classes (*High*, *Moderate* and *Low*), while no statistical significance was observed between the two moderate sub-classes (**Table A2.5**). Finally, hypocotyl and green berry assays were independently analyzed, and similar ranges for each class were found, with slight differences according to the plant material under-study (**Table**

A2.2), being the border limits of each class more difficult to define on green berries than in hypocotyls.

Table 3.2 - Detailed data description, for each *C. kahawae* isolate, of all aggressiveness quantitative traits (average values from both assays of hypocotyls and detached green berries), and subsequent scoring into aggressiveness classes and sub-classes.

Isolates	Aggressiveness quantitative traits						Classes	Sub-Classes
	AUDPC	IDI_10dai	nr_days_R4	nr_days_50%	nr_days_100%	Latent_period		
Ang 29	77.07 ± 2.73	1.00 ± 0.0	7 ± 1.15	7 ± 1.15	9 ± 1.15	6 ± 0		
Zim 12	74.82 ± 1.82	1.00 ± 0.0	9 ± 1.00	9 ± 1.00	10 ± 0.00	6 ± 0	***High***	***High***
Uga 5	74.26 ± 1.14	1.00 ± 0.0	8± 0	8± 0	10 ± 0.00	6 ± 0		
Uga 1	65.29 ± 8.89	0.78 ± 0.23	13 ± 3.32	11 ± 2.50	17 ± 3.56	10 ± 4.95		
Que 84	63.63 ± 10.31	0.70 ± 0.29	13 ± 3.32	12 ± 2.45	16 ± 4.27	12 ± 2.12		
Cam 1	70.25 ± 6.41	0.87 ± 0.17	10 ± 2.06	10 ± 2.06	15 ± 2.31	7 ± 1.41		
Que 2	67.82 ± 9.32	0.82 ± 0.28	11 ± 2.99	10 ± 3.30	16 ± 2.99	8 ± 2.83		***High_Moderate***
Eti 20	66.41 ± 5.46	0.81 ± 0.14	12 ± 2.45	12 ± 2.45	16 ± 2.99	8 ± 0		
Cam 2	68.72 ± 2.13	0.90 ± 0.04	11 ± 1.50	11 ± 1.50	16 ± 1.00	6 ± 0		
Que 42	62.09 ± 6.84	0.66 ± 0.0	13 ± 2.36	12 ± 1.50	17 ± 2.87	9 ± 1.41	***Moderate***	
Bur 2	65.23 ± 5.12	0.75 ± 0.17	13 ± 2.06	13 ± 2.06	18 ± 2.45	10 ± 4.95		
Rua 1	64.40 ± 4.97	0.75 ± 0.10	16 ± 4.27	14 ± 4.24	17 ± 4.73	10 ± 0		
Que 72	59.31 ± 7.61	0.63 ± 0.24	15 ± 3.30	14 ± 2.99	22 ± 1.91	9 ± 1.41		
Ang 6	56.68 ± 10.50	0.59 ± 0.30	18 ± 3.32	16 ± 4.79	21 ± 1.91	12 ± 2.12		***Low_Moderate***
Que 48	59.00 ± 3.38	0.63 ± 0.12	15 ± 1.00	14 ± 1.15	21 ± 2.00	9 ± 1.41		
Zim 1	62.52 ± 4.61	0.73 ± 0.15	15 ± 4.20	14 ± 4.24	17 ± 5.19	8 ± 0		
Que 82	46.08 ± 12.43	0.38 ± 0.21	(18 – >24) ± a	(15 – >24) ± a	(19 – >24) ± a	13 ± 3.54		
Tan 12	49.57 ± 7.66	0.44 ± 0.20	(18 – >24) ± a	19 ± 3.92	(22 - >24) ± a	8 ± 0		
Tan 13	36.07± 9.71	0.33 ± 0.13	(24 - >24) ± a	(22 – >24) ± a	>24 ± a	15 ± 7.07		
Mal 2	56.06 ± 6.18	0.53 ± 0.15	17 ± 2.36	15 ± 2.31	21 ± 3.40	13 ± 0		
Ang 67	51.28 ± 6.08	0.44 ± 0.12	17 ± 2.06	17 ± 2.06	22 ± 3.30	8 ± 2.83	***Low***	***Low***
Eti 9	45.06 ± 9.93	0.40 ± 0.16	(21 - >24) ± a	(19 – >24) ± a	>24 ± 0	12 ± 2.12		
Uga 3	54.60 ± 5.52	0.53 ± 0.05	(17 - >24) ± a	(15 – >24) ± a	(21 – >24) ± a	13 ± 0		
Zim 14	49.89 ± 4.64	0.46 ± 0.08	(17 - >24) ± a	(17 – >24) ± a	(21 - >24) ± a	9 ± 1.41		
Cam 8	55.27 ± 6.54	0.51 ± 0.16	(17 - >24) ± a	(16 – >24) ± a	(22 - >24) ± a	12 ± 2.12		
Cam 5	53.85 ± 6.47	0.57 ± 0.17	(20 - >24) ± a	20 ±3.69	(20 - >24) ± a	9 ± 1.41		

a* not computed

Moreover, some incongruence on isolate class attribution, was observed depending on the testing material, particularly within the two moderate sub-classes (**Figure A2.1**). For instance, Uga 1, Que 84, Rua 1, Que 42 and Zim 1 were always considered as moderate aggressive isolates switching between the two sub-classes according to the coffee organ tested. Mal 2 and Uga 3 were classified as low aggressive isolates in the hypocotyls and as moderate low aggressive isolates in green berries, while Ang 6 and Cam 8 had an inverted classification.

3.4.3.2 Cytological traits associated with agressiveness classes

Representative isolates of each aggressiveness class (*High*, *Moderate* and *Low*) were used for a cytological analysis. Firstly, a high percentage of germination (above 80%) and a high percentage of melanized appressoria (above 90%) were observed. Secondly, a high correlation was observed between the aggressiveness class of the isolates and its pattern of development during host infection. The lowest aggressive isolate, when compared with the highest and moderate aggressive isolates had a delay (around 1 dai) in the development of key infection stages (host tissue penetration, establishment of biotrophy, and switch to necrotrophy). Host tissue penetration was observed at 2 dai for both high and moderate aggressive isolates, but the hyphal length inside the host tissue was significantly higher in the most aggressive isolate, both at 2 dai and 3 dai (**Table 3.3** and **Figure 3.4**).

Table 3.3 - Evaluation of fungal growth as a measure of hyphal length in coffee hypocotyls, after challenge with *C. kahawae* isolates representative of high (Ang29), moderate (Que2) and low (Ang67) aggressive patterns at different times after inoculation.

Days after inoculation (dai)	Aggressiveness profile		
	high (Ang29)	moderate (Que2)	low (Ang67)
	Mean lenght (µm) ± SD	Mean lenght (µm) ± SD	Mean lenght (µm) ± SD
1	0±0	0±0	0±0
2	16.39±3.98 a	6.46±3.49 b	0±0 c
3	40.59±8.42 a	21.77±8.3 b	5.75±2.21 c

a) (X±SD) mean±standard deviation

b) The values with different letters in the same row are significantly different (P<0.05; Least Significant Difference (LSD)

On the other hand, the low aggressive isolate was only able to penetrate host tissues at 3 dai, showing a hyphal length significantly lower than those of isolates from the high and moderate aggressiveness classes (**Table 3.3** and **Figure 3.4**). Subsequently, the switch to necrotrophy was observed at 3 dai for the high and moderate aggressive isolates, and at 4 dai for the low aggressive isolate. Finally, the incubation period was of 3 dai for the higher and moderate aggressive isolates, and around 4 dai for the lower aggressive isolate. Despite this similar incubation period for high and moderate aggressive isolates, the number of R1 lesions observed was significantly higher in the most aggressive isolate (data not shown).

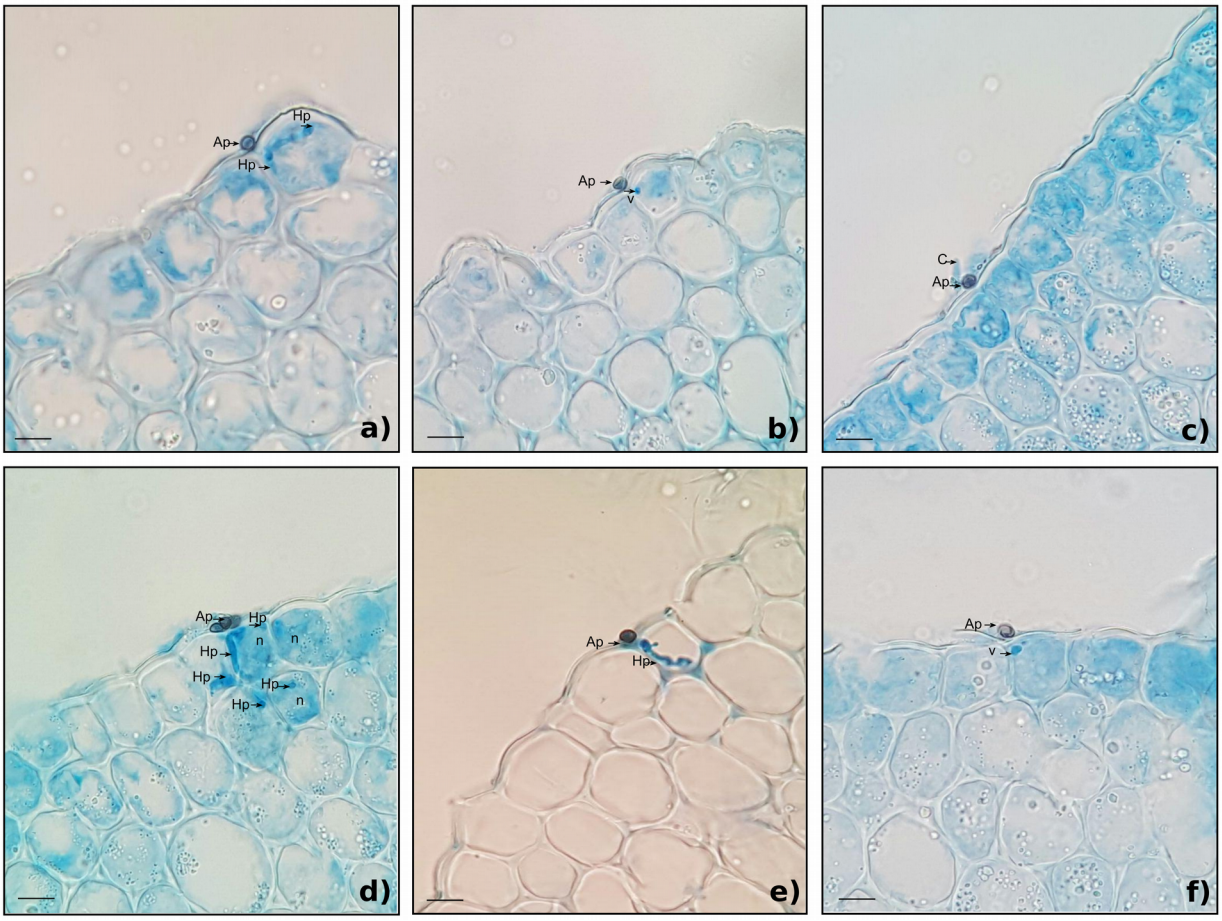


Figure 3.4 - *C. kahawae* post-penetration development in coffee susceptible hypocotyls (var. Caturra) of three aggressiveness representative isolates (Ang 29 – high aggressive isolate; Que 2 – moderate aggressive isolate and Ang 67 – low aggressive isolate). Light microscope observations with cotton blue lactophenol staining. Scale Bar= 10µm **a)** Infection site showing a melanized appressorium (Ap) and intracellular hyphae (Hp) of Ang 29 in the epidermal plant cell at 2 dai; **b)** Infection site showing a melanized appressorium (Ap) and an infection vesicle (v) of Que 2 in the epidermal plant cell at 2 dai; **c)** Conidium (C) and a melanized apressorium (Ap) of Ang 67 at 2 dai; **d)** Infection site showing a melanized appressorium (Ap) and intra- and intercellular hyphae (Hp) of Ang 29 in living and necrotized (n) host cells at 3 dai; **e)** Infection site showing a melanized appressorium (Ap) and an intracellular hypha(Hp) of Que 2 in a living epidermal plant cell, at 3 dai; **f)** Infection site showing a melanized appressorium (Ap) and an infection vesicle (v) of Ang 67 in a living epidermal plant cell at 3 dai

3.5 Discussion

In this work, a comprehensive characterization of *C. kahawae* aggressiveness was performed using complementary data obtained from a broad set of parameters. The *AUDPC* was shown to be a good indicator of aggressiveness and is one of the most used aggressiveness traits in the literature (Muiru *et al.*, 2010; Frézal *et al.*, 2012; Purahong *et al.*, 2012, 2014; Suffert *et al.*, 2013; Pires *et al.,* 2016), as it gives a reflection of the area of the organ covered by the disease. Nonetheless, this metric by itself is not enough to accurately characterize the isolates´aggressiveness, since it cannot directly account for the shape of the symptoms evolution curve, and therefore, additional quantitative traits need to be addressed. The *IDI_10dai* is a quantitative trait that is easier to measure (only one time-point) and, due to its high correlation with *AUDPC*, it can be considered a highly efficient aggressiveness trait that could be applied on future studies. The latent period is also one of the most used aggressiveness quantitative traits on the literature, and provides important information regarding the isolates adaptation. Since isolates with lower latent period have a selective advantage in the field, by quickly releasing the conidia into the environment (Pariaud *et al*., 2009a; Frézal *et al.*, 2012; Delmas *et al*., 2016). Besides that, the latent period is sometimes considered identical to incubation period (Luo & TeBeest, 1997). However, in coffee - *C. kahawae* interaction, as also described for *Dioscorea alata - C. gloeosporioides interaction* (Frézal *et al.*, 2012), conidia dispersion was not correlated with the emergence of first symptoms. Therefore, as suggested by our results and in accordance with Frezal *et al.* (2012), the incubation period may not be a relevant aggressiveness trait within the *C. gloeosporioides* species complex. The remaining quantitative traits were crucial to assess the time interval that takes to reach the main time-points of the infection, according to the isolates aggressiveness profile. Similar approaches were applied to select the best set of aggressiveness quantitative traits for other pathogens, identifying, for instance, the incubation and latent period, along with the development rate of sporulating area, maximal sporulating area, pycnidial density, and sporulation capacity as the most suitable quantitative traits for *Mycosphaerella graminicola* (Suffert *et al*., 2013), while for *Exserohilum turcicum* (Muiru *et al*., 2010) the incubation period, *AUDPC* along with size of chlorotic and necrotic lesions, lesion density and rate of

lesion expansion were considered the best aggressiveness parameters (Muiru *et al.*, 2010; Suffert *et al.*, 2013). Overall, aggressiveness is a very sensitive trait that often varies between isolates within a species and between species, and thus needs to be evaluated to each particular interaction and plant material, ideally with the resource of a comprehensive approach.

Another crucial goal of this work was to assess if the use of green berries and hypocotyls as an experimental system for testing aggressiveness was significantly different. For that, several correlation coefficient analyses were performed. Globally, the values of correlation coefficient obtained per isolate suggest a high assay reproducibility within isolates for each plant organ (hypocotyls and green berries). Furthermore, per inoculated plant material a high and a moderate correlation was observed for hypocotyls and green berries, respectively. These results suggest that hypocotyls provide a higher data reproducibility than green berries, which may be justified by a higher physiological homogeneity within hypocotyls, since it is easier to collect all plantlets in a similar developmental stage, than it is for green berries. In fact, although all green berries were collected in the expanding phase, they present slight differences and even the smallest developmental variations can influence the susceptibility response (Pinard *et al.*, 2012). Previous field studies showed that the susceptibility of green berries to *C. kahawae* varies with the developmental stage, being susceptible between (4-6 weeks) to (16-18 weeks) and resistance in the remaining period (Pinard *et al.*, 2012) . By contrast, the susceptibility of hypocotyls seems to be very similar within 2 -8 weeks (Murakaru, 1976). Finally, the correlation coefficient analysis between green berries and hypocotyls showed a high correlation coefficient value (r=0.77). This result suggests that the information retrieved from green berries and hypocotyls is correlated and could be equally used to perform the evaluation of an isolates aggressiveness. However, it is important to note that, previous studies with the purpose of testing resistant material showed a weak correlation coefficient value between these two coffee organs (r=0.24). This was possibly due to the loss of resistance when the green berries were detached from the plant (van der Vossen *et al.*, 1976). Nevertheless, in the context of an aggressiveness study, both green berries and hypocotyls, seems to be a reliable testing material to use in further studies, being hypocotyl a more reproducible material.

In this study, three discrete aggressiveness classes (*High*, *Moderate* and *Low*) were established, with the moderate class being sub-divided into two sub-classes: *High_Moderate* and *Low_Moderate*. Although some degree of cline variation was found across all aggressiveness traits studied, it was possible to group the isolates according to their profile. In general, the three main classes are defined by a set of standard metrics fairly constant between isolates, which will allow the assignment of different isolates with confidence. Except, perhaps, for the moderate sub-classes where the parameter values could be either stable or vary in some degree, but always within the ranges of the moderate ranking. Using this system, *C. kahawae* isolates were assigned into different classes and sub-classes and in generally presented a similar classification to those previously attributed in other *C. kahawae* aggressiveness studies (Loureiro *et al.*, 2011; Pires *et al.*, 2016; Vieira *et al.*, 2016). Nevertheless, a few discrepancies were found for some isolates, such as Cam 1 which was classified by Loureiro *et al.* (2011) as highly aggressive, and such as Que 2, Ang 6 and Zim 1 which were classified by Pires *et al.* (2016) as low aggressive isolates, as well as Zim 12 as moderately aggressive. In our study, these isolates revealed a different profile: Zim 1 and Ang 6 were classified as *Low_Moderate* aggressive isolates, Que 2 and Cam 1 as *High_Moderate* aggressive isolate and Zim 12 as a high aggressive isolate. These differences could be due to the absence of truly low aggressive isolates in the Pires *et al.*, (2016) study and truly high aggressive isolates in those described by Loureiro *et al.* (2011), or to the different metrics applied, leading to a slight lag in the classification scale. Since the establishment of aggressiveness classes directly depends on the metrics and the quantitative traits used, the lack of a comprehensive representation of aggressiveness profiles could lead to biases in isolate classification. Moreover, it is well known that aggressiveness is a very sensitive trait that can change according to the physiological state of the pathogen (storage/ multiplication conditions/age), inoculation, temperature, host physiology and plant material (Pariaud *et al.*, 2009a) and that can explain, at least in part, some of these incongruities. In addition, discrepancies in isolate classification were also observed depending on the aggressiveness assay system applied. Most of these incongruities occurred within the two moderate sub-classes, but three isolates (Ang 6, Mal 2 and Ug3) swap between two main aggressiveness class (*Moderate* and *Low*). In fact, Ang 6 seems to have an identical classification on green

berries to that reported by Pires *et al.*, (2016). These slight differences were expected not only because the isolate aggressiveness could be influenced by organ physiology, but also due to the continuum of variation observed that make the class definition somewhat difficult. Overall, these classes seem to be able to accommodate all the aggressiveness variation observed in *C. kahawae*.

A cytological analysis of representative isolates of each aggressiveness classes was made to evaluate differences in the pre- and post-penetration fungal growth stages. Overall, our results showed that *C. kahawae* aggressiveness can be related with the development of post-penetration stages, rather than conidia germination and appressoria differentiation. This analysis allowed the identification of the infection stages (fungal penetration, establishment of biotrophy, and switch to necrotrophy) from which it is possible to differentiate the isolates aggressiveness, namely, the lowest aggressive isolate has a delay (around 1dai) in the development of key infection stages, while moderate and high aggressive isolates penetrate the host tissue and switches to necrotrophy at the same time, but the hyphal length inside the host tissue is at least two times higher in the most aggressive isolate. In fact, it is quite evident that host colonization is much faster and invasive with the increase of aggressiveness, which strongly corroborates the results collected by the remaining quantitative traits. The cytological observations also provided evidences of why the incubation period is not a good metric to characterize the isolate aggressiveness, since isolates from both classes (*High* and *Moderate*) presented symptoms at the same time, but with a distinct number of lesions. However, further ultrastructural observations and gene expression analyses will be required to better understand the mechanisms underlying aggressiveness.

In conclusion, we used a broad set of metrics and a vast sampling of *C. kahawae* isolates to establish aggressiveness classes, able to characterize other isolates, whether they are preserved in collections or recently collected from the field. In a time where the development of coffee resistant varieties is still one of the most sustainable control measures against CBD, this study is of the utmost importance. Selection of tester isolates representative of the global *C. kahawae* range of aggressiveness will allow the improvement of pre-screening resistance tests within breeding programs for producing more effective resistant varieties. In this sense, all breeding programs should

perform an aggressiveness profiling of the local isolates before starting screening coffee materials, to ensure that a comprehensive testing is made against all the aggressiveness profiles observed in the field, and consequently a more inclusive resistance may be found. Therefore, the present study provides data and knowledge useful to develop new and improve ongoing investigation lines on this plant pathogen and offers the opportunity to engage on future genotype-phenotype studies.

3.6 References

Alkimim, E.R., Caixeta, E.T., Sousa, T.V., Pereira, A.A., Carlos, A., Oliveira, B., Zambolim, L. and Sakiyama, N.S. (2017) Marker-assisted selection provides arabica coffee with genes from other *Coffea* species targeting on multiple resistance to rust and coffee berry disease. *Mol. Breed.***37**, 6.

Australia Group (2014) Australia Group Common Control List Handbook – *Volume II: Biological Weapons-Related Common Control Lists. Barton, Australia*: *The Australian Group.*

Batista, D., Silva, D.N., Vieira, A., *et al.* (2017) Legitimacy and implications of reducing *Colletotrichum kahawae* to subspecies in plant pathology. *Front. Plant Sci.* **7**, 1–4.

Bedimo, J.A.M., Bieysse, D., Nyasse, S., Nottéghem, J.L. and Cilas, C. (2010) Role of rainfall in the development of coffee berry disease in *Coffea arabica* caused by *Colletotrichum kahawae*, in Cameroon. *Plant Pathol.* **59**, 324–329.

Beynon, S.M., Coddington, B. and Varzea, V. (1995) Genetic variation in the coffee berry disease pathogen, *Colletotrichum kahawae*. *Physiol. Mol. Plant Pathol.* **46**, 457–470.

Boedo, C., Benichou, S., Berruyer, R., *et al*. (2012) Evaluating aggressiveness and host range of *Alternaria dauci* in a controlled environment. *Plant Pathol.* **61**, 63–75.

Bridge, P.D., Waller, J.M., Davies, D. and Buddie, A.G. (2008) Variability of *Colletotrichum kahawae* in relation to other *Colletotrichum* species from tropical perennial crops and the development of diagnostic techniques. *J. Phytopathol.* **156**, 274–280.

Castiblanco, V., Castillo, H. and Miedaner, T. (2018) Candidate genes for aggressiveness in a natural *Fusarium culmorum* population greatly differ between wheat and rye head blight. *J. Fungi* **4**, 14.

Delmas, E.L.C., Fabre, F., Jérôme, J., Mazet, I.D., Cervera, S.R., Laurent, D. and François, D. (2016) Adaptation of a plant pathogen to partial host resistance:

selection for greater aggressiveness in grapevine downy mildew. *Evol. Appl.* **9**, 709–725.

Derso, E. and Waller, J.M. (2003) Variation among *Colletotrichum* isolates from diseased coffee berries in Ethiopia. *Crop Prot.* **22**, 561–565.

Figueiredo, A., Loureiro, A., Batista, D., Monteiro, F., Várzea, V., Pais, M.S., Gichuru, E.K. and Silva, M.C. (2013) Validation of reference genes for normalization of qPCR gene expression data from *Coffea* spp. hypocotyls inoculated with *Colletotrichum kahawae*. *BMC Res. Notes* **6**, 388.

Frézal, L., Desquilbet, L., Jacqua, G. and Neema, C. (2012) Quantification of the aggressiveness of a foliar pathogen, *Colletotrichum gloeosporioides*, responsible for water yam (*Dioscorea alata*) anthracnose. *Eur. J. Plant Pathol.* **134**, 267–279.

Gichuru, E.K., Agwanda, C.O., Combes, M.C., Mutitu, E.W., Ngugi, E.C.K., Bertrand, B. and Lashermes, P. (2008) Identification of molecular markers linked to a gene conferring resistance to coffee berry disease (*Colletotrichum kahawae*) in *Coffea arabica*. *Plant Pathol.* **57**, 1117–1124.

Graff, V. der (1981) Selection of arabica coffee resistant to coffee berry disease in Ethiopia. *Dr. thesis. Wageningen, Netherlands*, 110.

Hindorf, H. and Omondi, C.O. (2011) A review of three major fungal diseases of *Coffea arabica* L. in the rainforests of Ethiopia and progress in breeding for resistance in Kenya. *J. Adv. Res.* **2**, 109–120.

Jayawardena, RS., Hyde, KD., Jeewon, R., Li, XH., Liu, M., Yan, JY., (2016) Mycosphere Essay 6: Why is it important to correctly name *Colletotrichum* species? *Mycosphere* **7**, 1076–92.

Lee, D.H., Roux, J., Wingfield, B.D. and Wingfield, M.J. (2015) Variation in growth rates and aggressiveness of naturally occurring self-fertile and self-sterile isolates of the wilt pathogen *Ceratocystis albifundus*. *Iran. J. plant Pathol.* **64**, 1103–1109.

Loureiro, A., Guerra-Guimarães, L., Lidon, F.C., Bertrand, B., Silva, M.C. and Várzea, V. (2011) Isoenzymatic characterization of *Colletotrichum kahawae* isolates with different levels of aggressiveness. *Trop. Plant Pathol.* **36**, 287–293.

Luo, Y. and TeBeest, D.O. (1997) Behavior of a wild-type and two mutant strains of *Colletotrichum gloeosporioides* f. sp. *aeschynomene* on northern jointvetch in the field. *Plant Dis.* **82**, 374–379.

Luzolo, M., Talhinhas, P., Várzea, V. and Neves-Martins, J. (2010) Characterization of *Colletotrichum Kahawae* isolates causing coffee berry disease in Angola. *J. Phytopathol.* **158**, 310–313.

Manga, B., Bieysse D, Mouen B J A, Akalay I, Bompard E, Berry D. (1998) Observation sur la diversité de la population de *Colletotrichum kahawae* agent de l'anthracnose des baies du cafeier Arabica. Implications pour l'am? eloration getique. *Proc. 17th Int. Conf. Coffee Sci. 1997, Nairobi, Kenya. Paris, Fr. Assoc. Sci. Int. du Cafe*, 604–12.

Muiru, W.., Koopmann, B., Tiedemann, A.., Mutitu, E.. and Kimenju, J.. (2010) Race typing and evaluation of agressiveness of *Exserohilum turcicum* isolates of Kenyan, German and Austrian origin. *World J. Agric. Sci.* **6**, 277–284.

Mulinge, K. (1970) Development of coffee berry disease in relation to the stage of berry growth. *Annals of Applied Biology* **65,** 269-27.

Murakaru, G. (1976) Influence of age of *Coffee* seedlings on infection by *Colletotrichum coffeanum* (Noack). *Kenya Coffee*, 55–57.

Omondi, C., Hindorf, H., Welz, H., Saucke, D., Ayiecho, P. and Mwang'ombe, A. (2000) Reaction of some *Coffea arabica* genotypes to strains of *Colletotrichum kahawae*, the cause of coffee berry disease. *J. Phytopathol.* **148**, 61–63.

Pariaud, B., Goyeau, H., Halkett, F., Robert, C. and Lannou, C. (2012) Variation in aggressiveness is detected among *Puccinia triticina* isolates of the same pathotype and clonal lineage in the adult plant stage. *Eur. J. Plant Pathol.* **134**, 733–743.

Pariaud, B., Ravigné, V., Halkett, F., Goyeau, H., Carlier, J. and Lannou, C. (2009) Aggressiveness and its role in the adaptation of plant pathogens. *Plant Pathol.* **58**, 409–424.

Pariaud, B., Robert, C., Goyeau, H. and Lannou, C. (2009) Aggressiveness components and adaptation to a host cultivar in wheat leaf rust. *Ecol. Epidemiol.* **99**, 869–878.

Pinard, F., Omondi, C.O. and Cilas, C. (2012) Detached berries inoculation for characterization of coffee resistance to coffee berry disease. *J. Plant Pathol.* **94**, 517–523.

Pires, A.S., Azinheira, H.G., Cabral, A., *et al*. (2016) Cytogenomic characterization of *Colletotrichum kahawae*, the causal agent of coffee berry disease, reveals diversity in minichromosome profiles and genome size expansion. *Plant Pathol.* **65**, 968–977.

Purahong, W., Alkadri, D., Nipoti, P., Pisi, A., Lemmens, M. and Prodi, A. (2012) Validation of a modified Petri-dish test to quantify aggressiveness of F*usarium graminearum* in durum wheat. *Eur. J. Plant Pathol.* **132**, 381–391.

Purahong, W., Nipoti, P., Pisi, A., Lemmens, M. and Prodi, A. (2014) Aggressiveness of different *Fusarium graminearum chemotypes* within a population from Northern-Central Italy. *Mycoscience* **55**, 63–69.

Rodrigues, C.J., Varzea, V.M. and Mederios, E.F. (1992) Evidence for the existence of physiological races of *Colletotrichum coffeanum* Noack sensu Hindorf. *Kenya Coffee (Kenya)* **57**, 1417–1420.

Silva, C., Várzea, V., Guerra-guimarães, L., Azinheira, H.G., Fernandez, D., Petitot, A., Bertrand, B., Lashermes, P. and Nicole, M. (2006) Coffee resistance to the main diseases : leaf rust and coffee berry disease. *Braz. J. Plant Physiol.* **18**, 119–147.

Silva, D.N., Talhinhas, P., Cai, L., Manuel, L., Gichuru, E.K., Loureiro, A., Várzea, V., Paulo, O.S. and Batista, D. (2012) Host-jump drives rapid and recent ecological speciation of the emergent fungal pathogen *Colletotrichum kahawae*. *Mol. Ecol.* **21**, 2655–2670.

Silva, M., Nicole, M., Rijo, L., Geiger, J. and Rodrigues Jr., C. (1999) Cytochemical aspects of the plant–rust fungus Interface during the compatible Interaction *Coffea arabica* (cv. Caturra) – *Hemileia vastatrix* (race III). *Int. J. Plant Sci.* **160**, 79–91.

Suffert, F., Sache, I. and Lannou, C. (2013) Assessment of quantitative traits of aggressiveness in *Mycosphaerella graminicola* on adult wheat plants. *Plant Pathol.* **62**, 1330–1341.

Várzea V.M.P., Rodrigues JCJ, Medeiros E (1993) Different pathogenicity of CBD isolates on coffee genotypes. *In: Proceedings of the 15th International Scientific Colloquium on Coffee, Montpellier, France. Paris, France: Association for Science and Information on Coffee. 303-308.*

Várzea, V.M.P, Rodrigues, J.C. and Lewis, B. (2002) Distinguishing characteristics and vegetative compatibility of *Colletotrichum kahawae* in comparison with other related species from coffee. *Plant Pathol.* **51**, 202–207.

Várzea, V.M.P, Rodrigues, C.J., Silva, M., Pedro, J. and Marques, D. (1999) High virulence of a *Colletotrichum kahawae* isolate from Cameroon as plant pathology compared with other isolates from other regions. *InProceedings 18th Int. Conf. Coffee Sci. 1999, Helsinki, Finland. Paris, Fr. Assoc. Sci. Int. du Cafe*, 131.

Vieira, A., Cabral, A., Fino, J., *et al.* (2016) Comparative validation of conventional and RNA-Seq data-derived reference genes for qPCR expression studies of *Colletotrichum kahawae*. *PLoS One* **11**, e0150651.

Vieira, A., Silva, DN., Várzea, V., Paulo, OS., Batista, D. (2018) Novel insights on colonization routes and evolutionary potential of *Colletotrichum kahawae*, a severe

pathogen of *Coffea arabica*. *Molecular Plant Pathology*. doi: 10.1111/mpp.12726.

Van der Vossen, H., (2009) The cup quality of disease-resistant cultivars of arabica coffee (*Coffea arabica*). *Experimental Agriculture* **45**, 323.

Vossen, H.A.M. van der, Cook, R.T.A. and Murakaru, G.N.W. (1976) Breeding for resistance to coffee berry disease caused by *Colletrotrichum caffeanum* Noack (sensu hindorf) in *Coffea arabica* L. I. Methods of preselection for resistance. *Euphytica* **25**, 733–745.

Genome-wide signatures of selection in *Colletotrichum kahawae* reveal candidate genes potentially involved in host specialization

4

Vieira A.[a,b,c], Silva DN.[a,b,c], Várzea V.[a,c], Paulo OS.[b], Batista D.[a,b,c]

[a]**CIFC/ISA** - UL, Oeiras, Portugal; [b]**CoBiG2/cE3c/FCUL** - UL, Lisboa, Portugal; [c]**LEAF/ISA -** UL, Lisboa, Portugal

4.1 Abstract

Plants and their pathogens are engaged in continuous evolutionary battles, with pathogens evolving to circumvent plant defense mechanisms and plants responding through enhanced protection to prevent or mitigate damage induced by pathogen attack. Artificial ecosystems are composed of genetically identical populations of crop plants with little changes from year to year. These environments are highly conducive to the emergence and dissemination of pathogens and they exert selective pressure for both quantitative traits linked to pathogen fitness, such as aggressiveness, and qualitative virulence factors responsible for fungi pathogenicity. In this study, we used a comparative genome-wide approach to investigate the genomic basis underlying the pathogenicity and aggressiveness of the fungal plant pathogen *Colletotrichum kahawae* infecting green coffee berries. The pathogenicity was investigated by comparing genomic variation between *C. kahawae* and its non-pathogenic sibling species, while the aggressiveness was studied by a genome-wide association approach with groups of isolates with different phenotypic profiles. High genetic differentiation was observed between *C. kahawae* and the most closely related species with 5 560 diagnostic SNPs identified, in which a significant enrichment of non-synonymous mutations was detected. Functional annotation of these non-synonymous mutations revealed a significant enrichment mainly in two gene ontology categories, “oxidation-reduction process” and “integral component of membrane”. Finally, the annotation of several genes potentially under-selection revealed that *C. kahawae’s* pathogenicity, may be a complex biological process in which important biological functions such as, detoxification and transport, regulation of host and pathogen gene expression, and signaling are involved. On the other hand, the genome-wide association analyses for aggressiveness were able to identify 10 SNPs and 15 SNPs of small effect in single and multi-association analysis, respectively, from which 7 were common. The annotation of these genomic regions allowed the identification of four candidate genes (“F-box domain-containing”, “nitrosoguanidine resistance”, “Fungal specific transcription factor domain-containing” and “C6 transcription factor”) that could be associated with aggressiveness. This study provides, for the first time, lights on the mechanisms and loci putatively involved in *C. kahawae* aggressiveness and pathogenicity.

4.2 Introduction

Plant diseases have become one of the most challenging threats to modern agriculture, not only for their huge economic impact caused by severe production losses, but also due to a global food security problem. Fungi are among the most devastating plant pathogens given their ability to overcome plant defenses and exploit the host's resources for their own reproduction and dispersion (Möller and Stukenbrock, 2017). In fact, plants and their pathogens are involved in a continuous battle, with pathogens evolving to suppress plant defenses and plants responding through enhanced protection mechanisms to reduce or suppress the pathogen damage, leading to a co-evolutionary dynamics that shapes the genomic landscape of both plants and pathogens (Möller and Stukenbrock, 2017; Zhan *et al.*, 2014). In natural systems, this co-evolution is tempered by host and environmental heterogeneity as well as pathogen trade-offs between pathogenicity and several life style traits (Zhan *et al.*, 2014; Zhan *et al.*, 2015). By contrast, in managed ecosystems, crops evolve through artificial selection, in which agriculturally desired traits are favored and the genetic heterogeneity of the host is severely reduced (Möller and Stukenbrock, 2017). In such homogeneous environments, the pathogen has a selective advantage, and newly pathogenic strains can quickly increase in frequency and spread across the fields (Zhan *et al.*, 2014). The genetic homogeneity of these environments also means that pathogens spend more time in a single selective environment when compared to the wild system. Therefore, it is likely that the host exert a selective pressure for quantitative traits linked to pathogen fitness, such as aggressiveness, as they do for qualitative virulence factors responsible for fungi pathogenicity (Elad and Pertot, 2014). Currently, the majority of plant pathogen studies are focused on the ability of the pathogen to infect the host (pathogenicity), and only few studies have focused their attention on the quantitative aspects of host-pathogen interaction (aggressiveness). However, it has been argued that it is the combination of these two approaches that will guide the formulation of sustainable disease management strategies, that can minimize disease epidemics while simultaneously reducing pressure on pathogens to evolve and increase in pathogenicity and aggressiveness (Pariaud *et al.*, 2009; Zhan *et al.*, 2014; Zhan *et al.*, 2015). From an evolutionary perspective, it is well known that the host is the strongest driver of

pathogen evolution, as a successful infection is required for pathogen reproduction and dispersal. In this sense, genes related to pathogenicity are expected to be under strong selective pressure, and consequently, genomic signatures of selection can be used to identify candidate genes involved in host-pathogen interactions (Möller and Stukenbrock, 2017). However, the potential of pathogens to evolve in response to host selective pressures can also be constrained by trade-offs in quantitative traits, namely the rate of infection progression. In fact, the existence of phenotypic variation in aggressiveness is a key factor necessary for pathogen adaptation (Delmas *et al.*, 2016). Hence, aggressiveness can be assessed by evaluating multiple phenotypic quantitative traits of the pathogen directly linked to its fitness. These traits are likely to be also under selection, resulting in differential adaptive patterns according to the environment (Pariaud *et al.*, 2009).

Nowadays, thanks to the development of High-throughput sequencing (HTS) a new era in plant pathology has emerged, making possible to unveil the genetic mechanisms underlying the pathogenicity and aggressiveness of pathogens (Byers *et al.*, 2016; Grünwald *et al.*, 2016). Genome scans for detecting genomic regions under positive selection can be used to identify genes involved in the adaptation, both within and between closely related species, while genome-wide association studies (GWAS) can identify genomic regions associated with a particular phenotype (Byers *et al.*, 2016; Grünwald *et al.*, 2016). Thus, a precise and reproducible measure of the relevant phenotype is the major limitation of GWA studies (Talas *et al.*, 2016). Both these approaches have been used in fungi to investigate host adaption (Connelly and Akey, 2012; Dalman *et al.*, 2013; Gao *et al.*, 2016; Palma-Guerrero *et al.*, 2013; Talas *et al.*, 2016), but their application is sill in their infancy compared to model plant and animal systems. Moreover, the genes identified as putatively under selection or associated with a phenotype are only candidates that require further experimental testing to determine how they affect the phenotype (Grünwald *et al.*, 2016).

Colletotrichum kahawae Waller & Bridge is a highly aggressive and specialized fungal pathogen, causing Coffee Berry Disease (CBD) in Arabica coffee in Africa. This pathogen emerged within the *C. gloeosporioides* complex, as a specialist pathogen with the ability to infect green coffee berries, an ecological niche previously unoccupied by

other fungi (Silva *et al.*, 2012). CBD can lead to severe production losses that reach up to 80% in extremely wet years, if no control measures are applied (Silva *et al.*, 2006), and, for that reason, *C. kahawae* was ranked as a quarantine pathogen and considered as a biological weapon (Batista *et al.*, 2017). Consequently, the pathogen´s potential dispersal to other Arabica coffee cultivation regions is greatly feared, particularly to those at higher altitudes in Latin America and Asia (Batista *et al.*, 2017). So far, no absolute effective control measure has been developed but some *Coffea* spp. genotypes show high levels of resistance (Várzea, VMP *et al.*, 2002). *C. kahawae* has also been described as a pathogen with a low genetic variability, clearly structured into three clonal and completely differentiated populations (Angolan, Cameroonian and East African) (Silva *et al.*, 2012), and two clonal lineages within the Angolan population (see chapter 2). Furthermore, significant differences in aggressiveness of isolates were consistently observed, regardless of their geographic origin (Bridge *et al.*, 2008; Loureiro *et al.*, 2011; Pires *et al.*, 2016), which led us to perform a comprehensive analysis and characterization of *C. kahawae* aggressiveness trait (see chapter 3). By providing consistent phenotypic data on aggressiveness, this study brought the opportunity to perform, for the first time, a GWAS for this trait in this pathogen. Up to now, and in contrast with other *Colletotrichum* species, little is still known about the adaptive genetic variation of *C. kahawae* and no reports have been made on candidate genes underlying its pathogenicity and/or aggressiveness. Therefore, the current work aims to: i) understand the genomic basis underlying the pathogenic behavior of *C. kahawae* on green coffee berries using a genomic comparative analysis with closely related non-pathogenic fungi, and ii) identify the genomic regions potentially associated with aggressiveness through a GWAS. These results will contribute to better understand the genomic basis underlying these two complex processes, which may allow the establishment of more evidence-based and effective control measures in the future.

4.3 Material and Methods

4.3.1 Sampling, DNA isolation and RAD - Sequencing

In this work, thirty *C. kahawae* isolates (CIFC/ISA/ULisboa collection) representative of the three genetic groups described by Silva *et al.*, (2012) and covering almost all

regions where the disease exists (ten African countries) were used, as well as ten isolates from non-pathogenic sibling species (**Table A3.1**). According to Weir *et al.*, (2012) the isolates belong to three different species (*Gomera cingulata* "f.sp.camelliae", *C. aotearoa* and *C. kahawae* subsp. *ciggaro*). However, in this study, the two *C. kahawae* subspecies sensu Weir *et al.*, (2012) are accepted as cryptic species as suggested by Batista *et al.*, (2017), and described accordingly. Therefore, our sampling comprises 5 isolates from *C. ciggaro*, 1 isolate from *G. cingulata* "f.sp.camelliae", and 4 isolates from *C. aotearoa*, all of them collected from different hosts and several countries across the world (**Table A3.1**). Culturing and DNA extraction from fungal isolates were performed as previously described by Silva *et al.*, (2012), with slight modifications. Briefly, isolates were grown in liquid media containing 3% malt extract and 0.5% peptone, under a photo-period of 12 h at 22°C. DNA was extracted from freeze dried mycelia with the Sigma Plant/Fungi DNA isolation kit (Sigma–Aldrich, Darmstadt, Germany), according to the manufacturer's instructions. Genomic DNA quality was evaluated by agarose gel and quantified using a Thermo Scientific (Waltham, MA, USA) Nanodrop ND-1000 spectrophotometer.

Three micrograms of high quality genomic DNA per sample were sent to Floragenex Inc. (Oregon USA) for RAD library preparation and sequencing. Libraries with sample-specific barcode [8 nucleotide (nt)] sequences were produced from DNA digested with PstI. RAD-seq pools were 100bp single-end-sequenced in a lane of an Illumina HiSeq 2000 machine. The sequence data was deposited in the European Nucleotide Archive under submission number PRJEB26929 and PRJEB28813.

4.3.2 RADseq quality filtering and SNP calling

Sequence reads were de-multiplexed and quality filtered with the *process_radtags* program from the package Stacks v1.20 (Catchen *et al.*, 2013). Reads with uncalled bases or distance to barcodes higher than 1 were removed. Base calls with a Phred score under 20 were converted to Ns and reads containing more than 4 Ns were discarded. Barcodes and Illumina adapters were excluded from each read and length was truncated to 85bp (-t 85). Additional filtering, and *de novo* assembly within and between individuals to identify loci was performed using the program PyRAD v3.0.5 (Eaton, 2014). This software was chosen due to its ability to handle indels when

clustering sequence reads into orthologous loci. In this study several clustering parameters were tested in order to minimize the number of missing data and maximize the number of phylogenetic informative sites (**Table A3.2**). The sequence variants [single nucleotide polymorphisms (SNPs)] were then exported into a variant call format (VCF) and the "stacks" information exported as a loci file. Handling and exploration of alignment data matrices was performed using TriFusion v1.0.0 software (https://github.com/OdiogoSilva/TriFusion).

4.3.3 Phylogenetic analysis

To assess phylogenetic relationships among the isolates we used a single concatenated alignment that includes loci with SNPs represented in more than 80% of the isolates and a minor allele frequency above 5% (*total_dataset*). Concatenation and conversion of the alignment matrices to the appropriate formats was performed with TriFusion. A maximum likelihood analysis was conducted with RAxML v. 8. 2 (Stamatakis, 2014) on the CIPRES Portal (Miller *et al.*, 2010), using the general time-reversible (GTR) model of nucleotide substitution with the CAT distributed rate heterogeneity. Nonparametric bootstrapping was performed with the fast bootstrap algorithm of RAxML with 1000 replicates using the GTRCAT substitution model. Bayesian inference was performed using MrBayes v3.2.6 (Ronquist *et al.*, 2012) with the GTR + Γ model of sequence evolution. The best-fitting model was determined according to the Akaike information criterion (Posada and Buckley, 2004). Posterior probabilities were generated from 1 x 10^7 generations, sampling at every 1000^{th} iteration, and the analysis was replicated three times with one cold and three incrementally heated Metropolis-coupled Monte Carlo Markov chains, starting from random trees. The achievement of the stationary phase and mixing was checked for all parameters using Tracer V1.4, and 1 x 10^6 generations (corresponding to 10% of the total of generations) were discarded as burn-in. Trees from different runs were combined using Logcombiner and summarized in a majority rule 50% consensus tree. All trees were visualized in FigTree (http://tree.bio.ed.ac.uk/software/figtree/) and further edited in Inkscape (https://inkscape.org/pt/). Note that, regardless of the dataset under study (datasets generated with different PyRAD parameters) a similar phylogenetic tree was reconstructed.

4.3.4 Detection of genomic signatures of positive selection related to the pathogenicity of *C. kahawae*

In this study, pathogenic (*C. kahawae*) and non-pathogenic fungi (*G. cingulata* "f.sp. *camelliae*", *C. aotearoa* and *C. ciggaro*) to *Coffee arabica* were analyzed in order to better understand the pathogenicity of *C. kahawae*. The initial dataset named as *total_dataset* comprise all the genetic variation observed within all the species (**Figure 4.1.a)**. While a second dataset named filtered_dataset was constructed using the diagnostic SNPs, ie the SNPs completely differentiated between pathogenic and non pathogenic groups, which were selected with the following sequential filters: i) by calculating the distribution of SNPs Fst values using VCFTOOLS v0.1.14 (Danecek *et al.*, 2011) and Arlequin v3.5.2 (Excoffier and Lischer, 2010) and choosing the SNPs with a Fst value equal to 1; ii) by choosing the SNPs that were conserved across all *C. kahawae* isolates and completely differentiated from at least one of the non-pathogenic fungi (**Figure 4.1.b**).

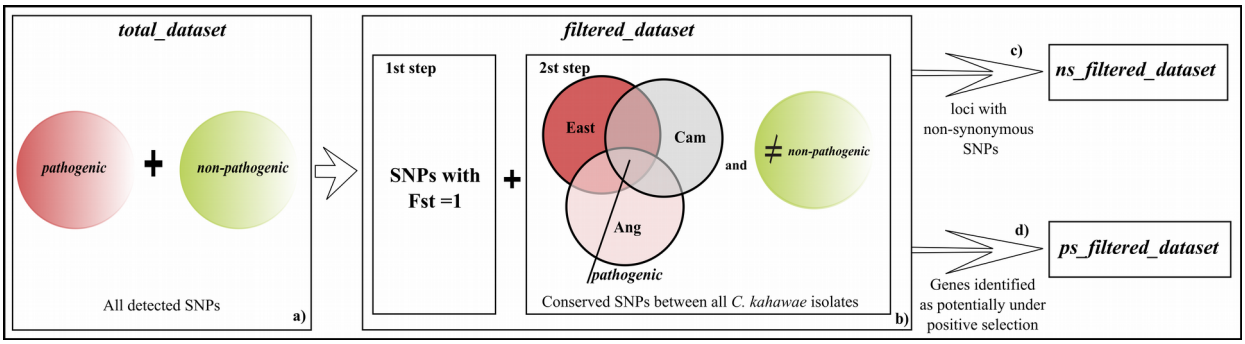


Figure 4.1 - Schematic representation of the datasets used for the analyses conducted in this study. **a)** *total_dataset* comprising all the detected SNPs; **b)** *filtered_dataset* comprising the diagnostic SNPs between pathogenic and non-pathogenic groups. The three *C. kahawae* populations were named as Ang (Angolan), Cam (Cameroonian) and East (East African). **c)** *ns_filtered_dataset* comprising all the loci with non-synonymous SNPs within the diagnostic SNPs; **d)** *ps_filtered_dataset comprising* all the genes potentially under positive selection

Both datasets, *filtered_dataset* and *total_dataset*, were mapped against the genome of the most closely related species within the *Colletotrichum* genus [*C. fruticola* (previously mis-identified as *C. gloeosporioides* Nara gc5 (Baroncelli *et al.*, 2016), accession_number (GCA_000319635.1) and reference (SAMN02981487)]. A copy of the assembled scaffolds was obtained from the Ensembl Genome Browser (useast.ensembl.org/index.html). All loci were then aligned to the reference genome using Bowtie 2.2.1.0 (Langmead and Salzberg, 2012) with the “--very-sensitive-local

default" setting. Alignments were sorted with SAMTools 0.1.19 (Li *et al.*, 2009) and the loci that aligned to more than one location were removed from the analysis. The SNPs location, annotation and classification of type of mutation were assessed with a custom-made python script available on https://github.com/yanavieira/Mapping_SNPs_Genome.git. The non-synonymous mutations identified in the *filtered_dataset* were used to create a new dataset named *ns_filtered_dataset*, containing only loci with non-synonymous SNPs (**Figure 4.1.c**). At this point, the consensus of the RADseq loci of the three datasets was functionally annotated. The categorization was made through a similarity BLASTx search using Blast2GO (Gotz *et al.*, 2008), against the NCBI non-redundant database with a minimum expectation value of 10^{-3}, and the remaining functional annotation was carried out using the default parameters. The Gene Ontology (GO) terms were assigned to the 2nd level of the biological process, molecular protein and cellular component categories. A GO enrichment analysis was performed to determine if any GO term was over or under represented in the *filtered_dataset* and *ns_filtered_dataset* when compared to the *total_dataset*. Statistically significant enrichment was tested against a reference of all genes analyzed using the Fisher's exact test and a significance of FDR< 0.05. Finally, the dN/dS ratio was measured for the genes identified in the *filtered_dataset*, and those having a ratio higher than 1 were considered as possible candidates under positive selection, and assemble as the *ps_filtered_dataset* (**Figure 4.1.d**). The annotation of these genes was further improved by searching with BLASTx the *C. kahawae* Rad loci (E-value ≤ 1e-1) and the orthologous *C. gloeosporioides* genes (E-value ≤ 1e-9) against the pathogen-host interaction reference database (PHI-base) v.4.2 (Urban *et al.*, 2017).

4.3.5 Genome wide association analysis for *C. kahawae* aggressiveness

The dataset used to perform the GWA study (*gwa_dataset*) was previously filtered in three steps to remove: i**)** all non-pathogenic fungi to green coffee berries; **ii)** all SNPs that contributed to the genetic structuring within *C. kahawae*, since the power of GWA can be significantly reduced by the inclusion of related individuals and population substructure (Connelly and Akey, 2012); **iii)** four isolates of *C. kahawae* that were not phenotypically classified in chapter 3.

The Bayesian Variable Selection Regression (BVSR) model proposed by Guan and Stephens, (2011) and implement in piMass v 0.9, was used to perform a single and multi-SNP correlation analysis between SNPs and the aggressiveness phenotype, using not only a pairwise comparative analysis between the three aggressiveness classes established in chapter 3 (*High*, *Moderate* and *Low*), but also a continuous analysis with the Area Under the Disease Progress Curve (*AUDPC*) parameter recorded for each isolate in chapter 3. An schematic representation of all the analyses performed and the datasets used is illustrated in **Figure 4.2**.

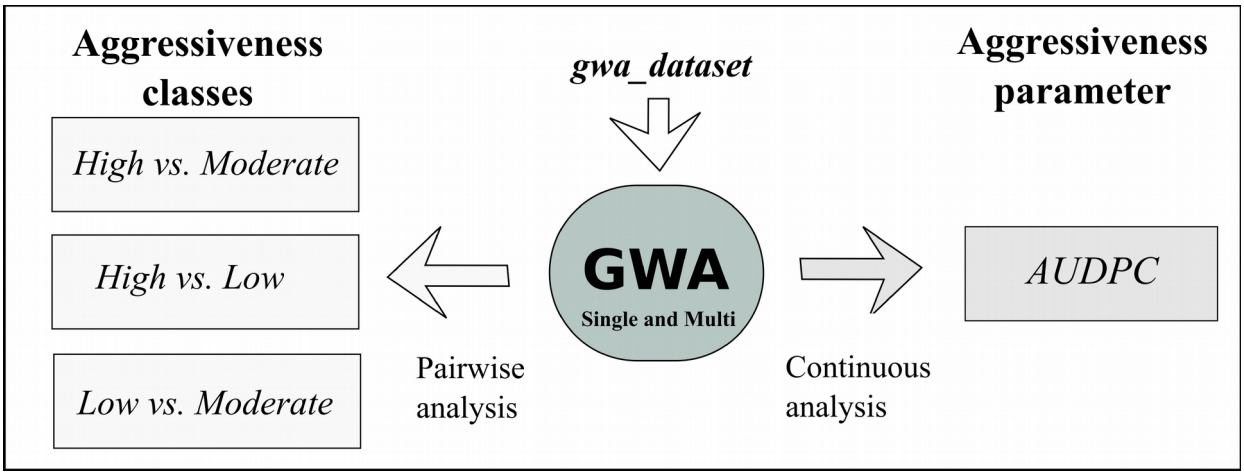


Figure 4.2 - Schematic representation of the dataset and GWA analyses conducted in this study. The pairwise analysis was performed taking into account the aggressiveness classes (*High*, *Moderate*, *Low*) previously described in chapter 3 and the continuous analysis was performed with the AUDPC values obtained in chapter 3.

This software was conceived to perform association studies with continuous response variables like *AUDPC*, but it is also appropriated for studies with binary phenotype data such as aggressiveness classes (Guan and Stephens, 2011). The BVSR method uses the phenotype as the response variable and the genetic variants (SNPs) as covariates to evaluate SNPs that may be associated with a particular phenotype (Guan and Stephens, 2011). SNPs statistically associated with phenotypic variation were identified by the posterior distribution of γ, or the posterior inclusion probability (PIP). In our association analyses, markers with a PIP greater than 97.5% empirical quartile (PIP 0.975 SNPs) were considered as highly associated with an aggressiveness class. For all 0.975 SNPs the respective PIP and the estimates of their phenotypic effect (β) are reported. A positive β in the pairwise X-Y aggressiveness class analysis means that the

frequency of the minor allele (MAF) is higher in the Y aggressiveness class and a negative β means that MAF is higher in the X aggressiveness class. Thus, to investigate the phenotypic effect size of each PIP0.975 SNP, the |β| was considered. Additional parameters contained in the model were estimated from the data: proportion of variance explained by the SNPs (PVE), the number of SNPs in the regression model (nSNPs) and the average of phenotypic effect of the SNP contained in the model (σSNP). For all pairwise and continuous analyses, we obtained 4 million Markov Chain Monte Carlo samples from the joint posterior probability distribution of model parameters (recording values every 400 iterations), and discarded the first 100,000 samples as burn-in. A single-SNP analysis was also performed to detect the associated SNPs even in the absence of interactions between them (Guan and Stephens, 2011), being the SNPs with an empirical quantile for Bayes Factor (BF) above 97.5% (BF0.975 SNPs) considered as strongly associated with isolates' aggressiveness. Imputation of missing genotypes was performed in BIMBAM v1.0 (Servin and Stephens, 2007), in which the state of a non-genotyped marker is inferred from the haplotype of the other individuals. The loci where the SNPs potentially associated with the aggressiveness trait are located, regardless the type of GWAS analysis, were functionally annotated as previously described in point 4.2.4, including the search on PHI-base for the SNPs located in coding regions.

4.4 Results

4.4.1 RAD tag generation and *de novo* assembly

Illumina RAD-seq of 30 *C. kahawae* isolates, collected from almost all coffee regions where CBD occurs, and 10 isolates from several closely related *C. gloeosporioides* species complex, generated an average of 3.76×10^6 reads per sample, amounting to a total of 150.41×10^6 of 85 bp single-end reads after barcode trimming. The individual read number ranged between 1.46×10^6 to 6.14×10^6, after an initial quality filter to remove the low quality reads, in which an average of 5.83×10^5 reads were discarded. Ten *de novo* assemblies were performed and the results are summarized in **Table A3.2**. The best *de novo* assembly, i.e. the one that minimizes the number of missing data and maximizes the number of phylogenetically informative sites, was obtained with the

following parameters: *minimum depth of coverage* of 10, *maximum number of low quality* of 4, *clustering threshold* of 0.90, *minimal taxon coverage* of 5 and *maximum shared heterozigosity* of 3. Additional filtering steps, including the removal of SNPs with less than 80% of the taxa represented and a minor allele frequency (MAF) lower than 5%, gave rise to the *total_dataset* with 83 528 SNPs across 28726 loci and 40 isolates.

4.4.2 Phylogenetic analysis

The phylogenetic analysis of the *total_dataset* produced a completely resolved evolutionary tree for the *C. gloeosporioides* complex species under study (**Figure 4.3**). Overall, a clear genetic differentiation was observed between the pathogenic and non pathogenic species. The branches were well supported in both analyses (Maximum Likelihood and Bayesian analyses) with all species being monophyletic, except for *C. ciggaro* that seems to be paraphyletic. In fact, two isolates (*ICMP_12953* and *Cg_432*) are more differentiated from the remaining *C. ciggaro* isolates and may even belong to a different species. The most differentiated species of the *C. gloeosporioides* complex under study was *C. aotearoa*. Finally, a geographical structuring within *C. kahawae,* similar to the one previously described in chapter 2, was observed, in which three well-supported populations (Angolan, Cameroonian and East African) and two clonal lineages within Angolan population are evident.

4.4.3 Genomic regions underlying the pathogenicity of *C. kahawae*

In this study, the isolates were sorted into two groups, pathogenic and non pathogenic, according to their ability to infect green coffee berries. The pathogenic group has all *C. kahawae* isolates (comprising 3 297 SNPs), while the non pathogenic group has isolates from the three closely related species (*G. cingulata* "f.sp. *camelliae*", *C. aotearoa* and *C. ciggaro*) with a total of 71 503 SNPs. The genetic variability between the two groups comprises 83 528 SNPs, which as previously referred, constitutes the *total_dataset*. The diagnostic SNPs (*filtered_dataset)* were chosen based on two sequential filtering steps.

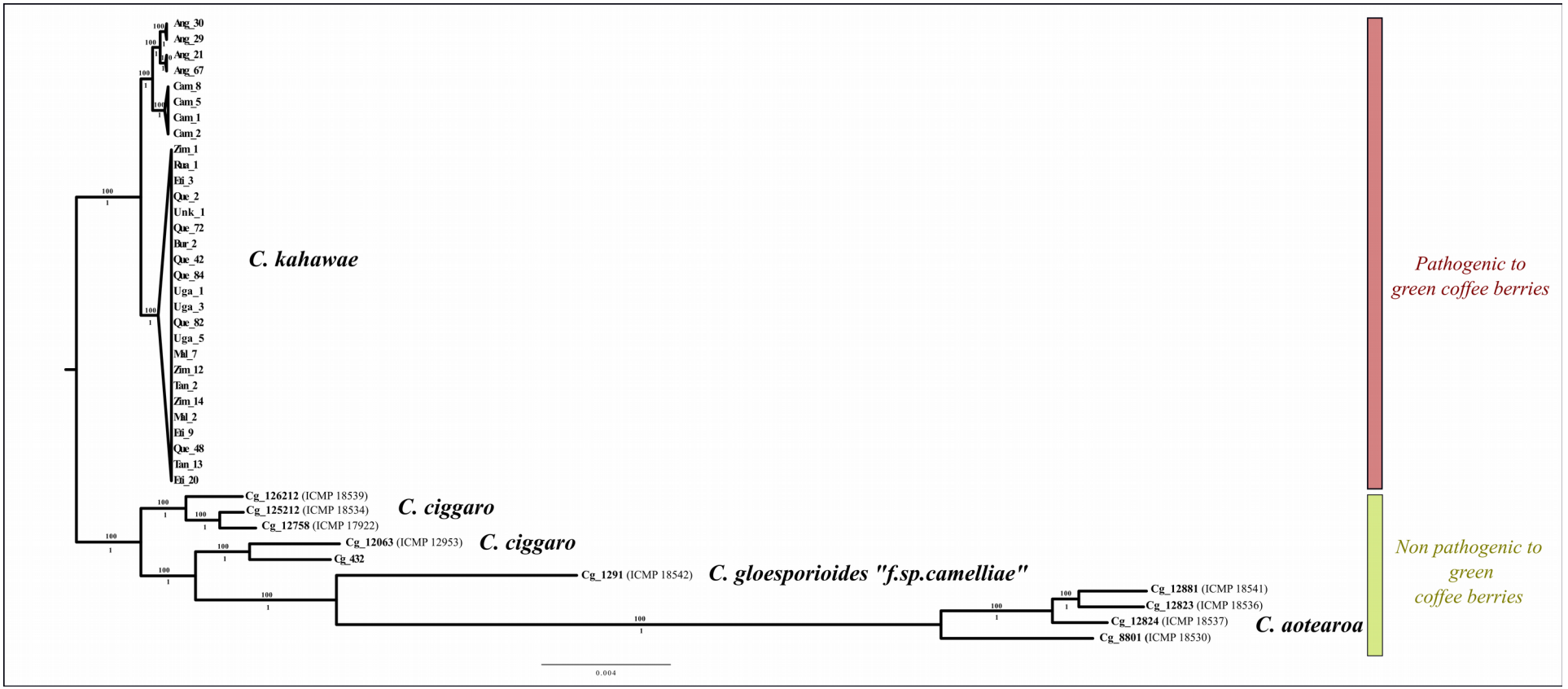


Figure 4.3 - Maximum likelihood phylogenetic tree illustrating the evolutionary relationships among pathogenic and non-pathogenic fungi to green coffee berries. Bootstrap and posterior probability values are provided above and below the branches.

The first filtering led to the identification of 7 773 SNPs located in 5 974 loci that are completely differentiated between the two groups (FST =1), while the second step reduced the *filtered_dataset* to a final group of 5 560 diagnostic SNPs located in 4 619 loci across 40 isolates.

Both datasets, *total_dataset* and *filtered_dataset*, were mapped against the genome of the most closely related species within the genus *Colletotrichum*, *C. fruticola* (Nara gc5). Only 28% (23 613 SNPs) of the *total_dataset* and 34 % (1 869 SNPs) of the *filtered_dataset* were successfully mapped. This analysis revealed that, in the *total_dataset*, 47% (11 162 SNPs) are located in non-coding regions, 53% (12 444 SNPs) are located in genes and 7 SNPs in pseudo genes, while in *filtered_dataset*, 55% (1 019 SNPs) are located in non-coding regions, 45% (847 SNPs) are located in genes and 3 SNPs are located in pseudo genes. Regarding the number of synonymous and non-synonymous mutations, a significant increase on the number of non-synonymous mutations was found in the *filtered_dataset* (45 %) when compared to the *total_dataset* (18%) (**Figure 4.4**). Therefore, a new dataset including only the loci with non-synonymous SNPs, and named as *ns_filtered_dataset*, was created to address this issue. The *ns_filtered_dataset* comprises 348 SNPs located in 336 loci.

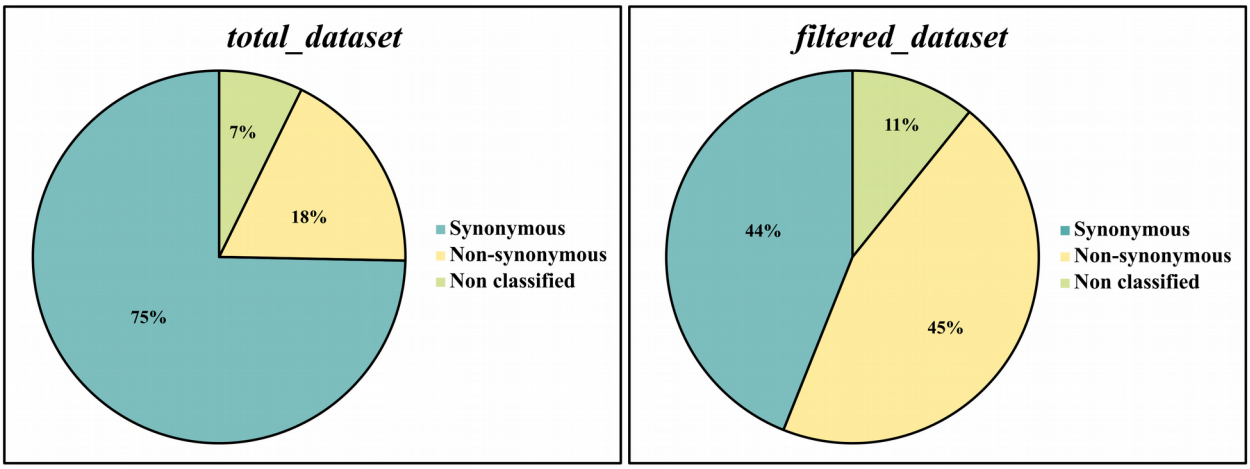


Figure 4.4 - Comparative analysis of the number of *synonymous* and *non-synonymous* SNPs in *total_dataset* and *filtered_dataset*

Functional annotation of RADseq loci of the three datasets (*total_dataset*, *filtered_dataset* and *ns_filtered_dataset)* was performed, in which 34%, 29% and 87% respectively, matched to known genes in the public nr database. GO terms were

assigned to 25% of the *total_dataset, 1*6% of the *filtered_dataset* and 57% of the *ns_filtered_dataset*, and analyzed at the 2nd level of functional annotation for biological process, molecular function and cellular component categories. Only small differences were observed between these three datasets (**Figure A3.1**). For the biological processes category, genes involved in "metabolic" and "cellular process" were highly represented in all datasets, while genes involved in "cellular component organization" were only present in the *total_dataset*. For the molecular functions category, "binding" and "catalytic activity" is the most represented GO term, being the "transporter activity" specific to both *filtered_dataset* and *ns_filtered_dataset*. For the cellular components category, the mostly represented functional classes in all datasets were "cell", "cell part", "membrane" and "membrane part". Additionally, fisher's exact test, between *ns_filtered_dataset* and *total_dataset*, revealed a significant enrichment of genes annotated into several GO terms, particularly for the functional classes *"*oxidation-reduction process" and "integral component of membrane*"* (**Figure 4.5**), while no significant differences were observed between the *filtered_dataset* and the *total_dataset*.

Finally, the analysis of dN/dS ratio on the *filtered_dataset* showed that 258 genes could be under positive selection (ratio > 1), from which 26 had more than 1 non-synonymous mutation. All genes potentially under selection were used to create the *ps_filtered_dataset* and its potential relationship with fungal virulence was searched on the pathogen-host interaction database (PHI-base), using two distinct approaches: i) blasting the complete gene retrieved from *C. fruticola* genome where the RAD loci mapped; ii) blasting the RAD loci of *C. kahawae* obtained during this study. A total of 77 *C. fruticola* genes had homology in the PHI-base, from which 40 agave also a match for *C. kahawae's* RAD loci (**Table A3.3**). From the total genes with a hit, 41 genes were reported to show an relevant role in fungal pathogenicity and virulence when a mutant phenotype was produced in other host-pathogen interactions (**Table A3.4**).

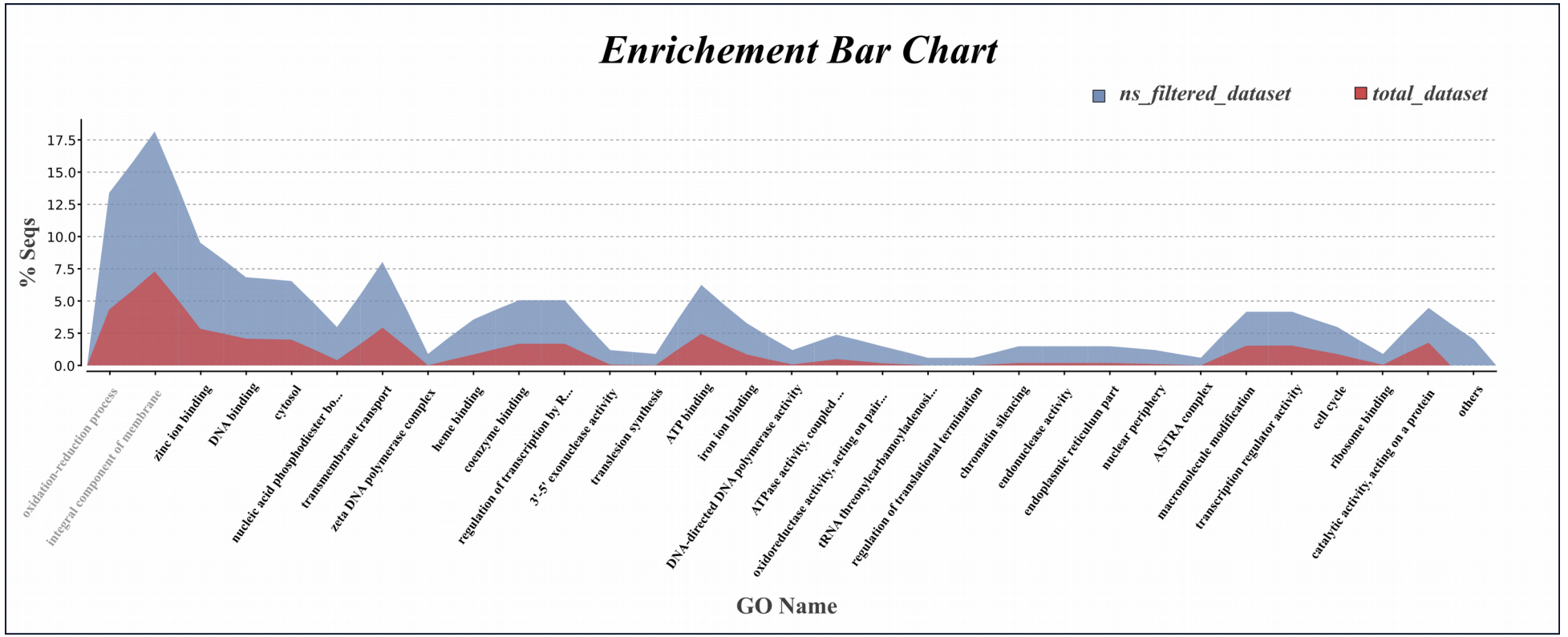


Figure 4.5 - Enrichment of gene functional categories among *total_dataset* and *ns_filtered_dataset.* Curve chart comparing the proportion of genes per GO term between *ns_filtered_dataset* and *total_dataset* with a statistically significance *of* FDR< 0.05, according to the Fisher's exact test. In light grey was evidenced the two most distinct GO term category.

The majority of these genes belonged to the category of “reduced virulence” in the PHI-base, while others belonged to different categories, including “loss of pathogenicity” (“chitin synthase”, “GTP-binding protein”, “ATP-binding cassette (ABC) transporter”, “alpha-mannosyltransferase cmt1” and “cytochrome p450”) and “lethal” (“ataxia telangiectasia mutated”, “C6 transcription factor”, “protein transport protein”, “ccr4-not transcription complex subunit”) (**Table A3.4**). Overall, the genes detected as being potentially under selection, and consequently, with a putative role in the pathogenicity of *C. kahawae* are mainly involved in oxidation-reduction processes and transport, but also genes involved in signaling, binding and biosynthesis processes are highly represented, which can have additional important roles in the infection process.

4.4.4 Genome-wide association study for the phenotypic trait of aggressiveness

After filtering the data for the GWAS, 173 SNPs located in 141 loci across 26 isolates, were identified and comprise the *gwa_dataset*. This dataset was used to tested for association with the phenotypic trait of aggressiveness. The efficiency of the filtering for correcting the effect of population genetic structure can be confirmed in **Figure A3.2**, showing that the selected SNPs were unable to recover the structuring pattern characteristic of *C. kahawae*.

The Single-SNP analysis, for all pairwise combinations (*High vs. Moderate*, *High vs. Low, Low vs. Moderate)* and continuous analysis (*AUDPC*), identified a total of 10 SNPs with BF0.975 (>97.5 quantile Bayes factor) associated with aggressiveness, corresponding to 6 % of the analyzed markers. When a more strict threshold was applied (99% quantile) 6 BF0.99 SNPs (3.5%) showed the strongest association with aggressiveness (41944.81; 34174.84; 18945.9; 18945.8; 12430.32; 46939.8). The number of SNPs identified in *High vs. Low* class was always 2 regardless the threshold used, while for the remaining pairwise analysis (*High vs. Moderate* and *Low vs. Moderate*) and continuous analysis (AUDPC) the number of SNPs range from 5 to 2 when a more restricted threshold was applied (**Figure 4.6 and Table 4.1**).

Table 4.1 - SNPs associated with aggressiveness for each pairwise comparison (*High vs. Moderate, High vs. Low, Low vs. Moderate)* and for the continuous analyses (*AUDPC*) obtained through Single-SNP association tests using Bayesian regression approach

SNP_ID	Alternative allele	Reference allele	SNP location	$BF_{0.975}$	| β |	Blast_hit	PHI-base (*Cf* gene)	Phi-base (*Ck* RADloci)
High vs. Low								
41944.81 [a]	G	T	CoR	0.08	0.06	F-box domain-containing	No hits	No hits
34174.84 [a,b]	C	T	CoR	0.08	0.06	nitrosoguanidine resistance	No hits	No hits
Mean_$BF_{0.975}$					0.06			
Mean all SNPs					-0.02			
High vs. Moderate								
18945.5	A	G	NcR	0.16	0.07	hypothetical protein	No hits	No hits
41944.81 [a]	G	T	CoR	0.16	0.07	F-box domain-containing	No hits	No hits
34174.84 [a,b]	C	T	CoR	0.28	0.08	nitrosoguanidine resistance	No hits	No hits
35951.85	T	C	CoR	0.14	0.06	Fungal specific transcription factor domain-containing	FZC28	GzZC278
Mean_$BF_{0.975}$					0.07			
Mean all SNPs					0			
Low vs. Moderate								
18945.9 [a,b]	C	T	NcR	0.18	0.09	hypothetical protein	No hits	No hits
14003.77 [b]	T	G	NcR	0.14	0.08	---NA--	x	No hits
46939.81 [a,b]	T	A	x	0.21	0.09	---NA--	x	No hits
7756.83	C	A	x	0.18	0.09	---NA--	x	No hits
Mean_$BF_{0.975}$					0.09			
Mean all SNPs					-0.02			
AUDPC								
18945.8 [a,b]	A	T	NcR	0.22	-1.86	hypothetical protein	No hits	No hits
18945.6 [b]	C	T	NcR	0.2	-1.8	hypothetical protein	No hits	No hits
12430.32 [a,b]	A	G	NcR	0.57	-2.49	---NA--	x	No hits
14003.77	T	G	NcR	0.19	-1.74	---NA--	x	No hits
34174.84	C	T	CoR	0.16	1.68	nitrosoguanidine resistance	No hits	No hits
Mean_$BF_{0.975}$					-1.24			
Mean all SNPs					0			

a- SNPs also selected with a $BF_{0.99}$; b- SNPs identified as potentially associated in single and multi-association analyses; CoR – coding Region; NCR – Non-coding region; x – No information; ---NA-- - No gene identified

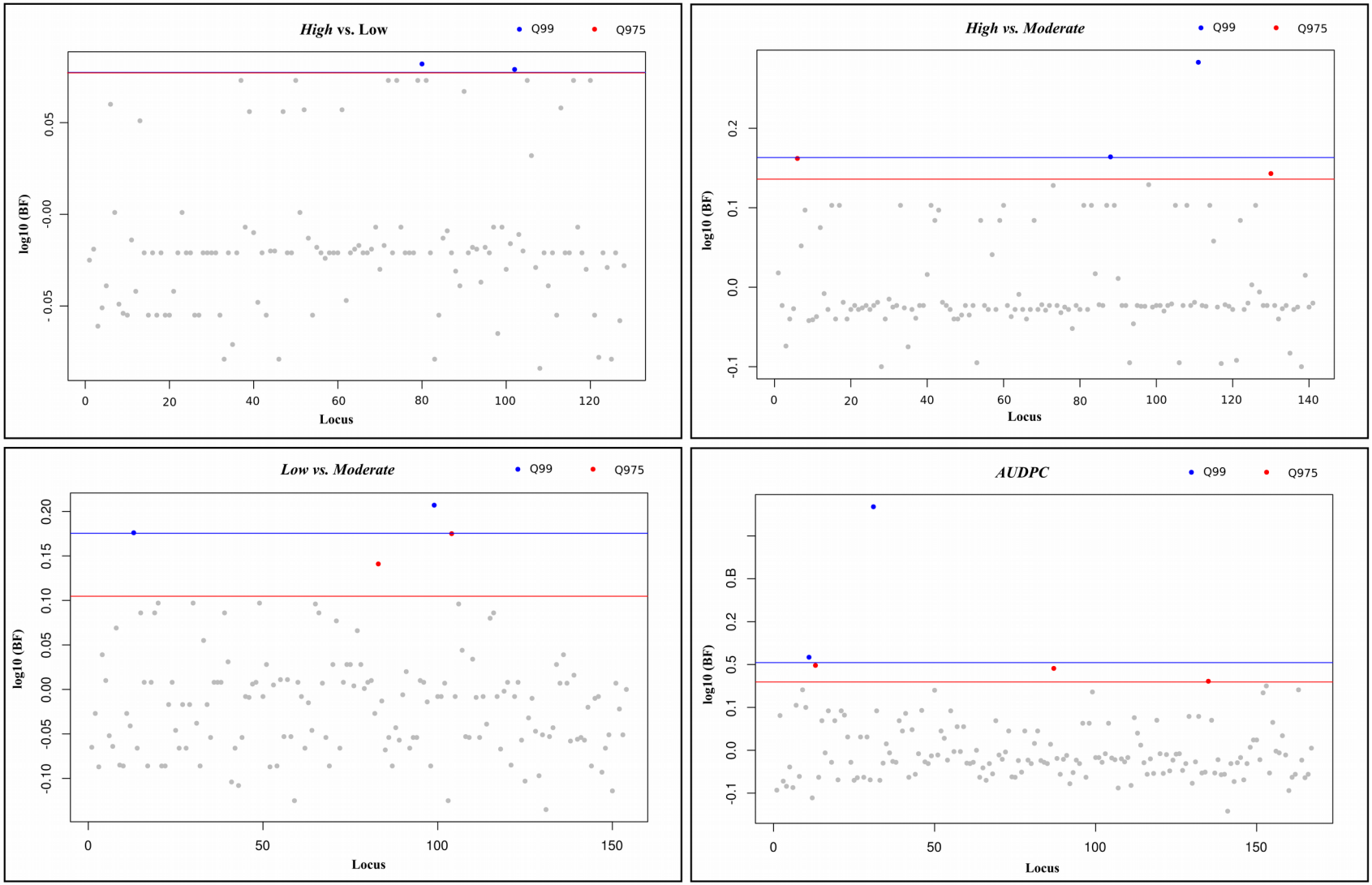


Figure 4.6 - Bayes factor for each analysis in Single-SNP association test. The horizontal blue lines correspond to the Bayes factor 99% empirical quantile threshold and red lines to the 97.5% empirical quantile. Blue dots: SNPs with a BF > 99% empirical quantile, Red dots: SNPs with a BF > 97.5% empirical quantile, Light grey dots: SNPs with a BF < 97.5% empirical quantile

For the multi-SNP association analysis, estimates of the mean number of SNPs (nSNPs) underlying the aggressiveness variation ranged from 26.5 to 52.2 SNPs (**Table 4.2**). When considering only models with the highest BFs (log10(BF) > BF0.99), the mean number of SNPs included in the model (nSNPs_BF) for each comparison decreased to values between 8.6 to 30.1, while the mean effect size of the SNPs (σSNP_BF) increased, ranging between 2.13 to 4.72 (**Table 4.2**). The posterior inclusion probabilities (PIP) for the analyzed SNPs were similar among all analysis but slightly higher in the pairwise comparisons involving *High vs. Moderate* (PIP = 0.314) classes and the continuous analyses with the *AUDPC* values (PIP= 0.312) (**Table 4.2**).

Table 4.2 - Parameter estimates from Bayesian variable selection regression for each pairwise analysis (*High vs. Moderate, High vs. Low, Low vs. Moderate)* and for the continuous analyses (*AUDPC*)

	PVE	σSNP	σSNP_BF	nSNP	nSNP_BF	PIP SNP
High vs. Low	0.442 (0.00-0.99)	0.766 (0.01 – 13.10)	4.080 (1.35 – 12.51)	33.5 (1.0 – 128.0)	8.6 (3 – 19)	0.261 (0.288)
High vs. Moderate	0.611 (0.00-0.99)	0.977 (0.01 – 15.03)	4.724 (1.35 – 13.22)	44.4 (1.0 – 141.0)	11.6 (5 – 22)	0.314 (0.357)
Low vs. Moderate	0.455 (0.00-0.99)	0.78 (0.01 – 44.37)	3.348 (0.99 – 9.72)	26.5 (1.0 – 154.0)	13.4 (4 – 34)	0.267 (0.346)
AUDPC	0.419 (0.00 – 0.99)	0.483 (0.00 – 9.391)	2.1360 (0.59 – 8.59)	52.2 (0.0 – 167.0)	30.1 (4 – 56)	0.312 (0.4037)

*PVE - proportion of variance explained by the SNPs; σSNP - average of phenotypic effect of the SNP that is in the model; σSNP_BF - average of phenotypic effect of the SNP that is in the model BF $_{0.99}$; nSNPs - the number of SNPs in the regression model; nSNPs_BF- the number of SNPs in the regression model BF $_{0.99;}$ posterior inclusion probability (PIP)

In multi-association analysis, a subset of 5 SNPs, with the highest inclusion probabilities (PIP0.975), were detected for all pairwise combinations (*High vs. Moderate*, *High vs. Low, Low vs. Moderate)* and continuous analysis (*AUDPC)* (**Figure 4.7**). Estimates of the strength of association between genotype variation at individual SNPs and phenotypic variation (|β|) were always greater than 0.36, but changed according to the analyses (**Table 4.3**). Overall, we obtained SNPs with larger effect in the continuous analysis than in the remaining pairwise analysis. Three PIP0.975 SNPs were shared between at least two pairwise analysis (34174.84; 18945.7; 14003.77), and no common SNP were detected between the continuous and all the pairwise analyses. In total, 15 different SNPs revealed a multi-SNP association with aggressiveness, from which 7 were also significant in the single-SNP analyses (34174.84; 18945.9; 14003.77; 46939.81; 18945.8; 18945.6; 12430.32). Despite that, no causal SNP, ie. able to explain the total phenotypic variation observed, was detected. The loci containing the

associated SNPs were BLASTed and functionally annotated, being described as genes located into “integral component of membrane” and “nucleus”.

Table 4.3 - SNPs associated with aggressiveness for each pairwise comparison (*High vs. Moderate, High vs. Low, Low vs. Moderate)* and for the continuous analyses (*AUDPC*) obtained through multi-SNP association tests using Bayesian regression approach

SNP_ID	Alternative allele	Reference allele	SNP location	PIP $_{0.975}$	| β |	Blast_hit	PHI-base (*Cf* gene)	Phi-base (*Ck* RADloci)
High vs. Low								
18638.64 [a]	G	A	x	0.3	0.61	---NA--	x	No hits
41138.71	G	A	NcR	0.28	0.37	---NA--	x	No hits
34174.84 [b]	C	T	CoR	0.28	0.43	nitrosoguanidine resistance	No hits	No hits
44503.84 [a]	A	G	CoR	0.29	0.45	C6 transcription factor	GzZC184	FZC55
High vs. Moderate								
18945.7	G	C	NcR	0.35	0.68	hypothetical protein	No hits	No hits
17838.69	C	T	CoR	0.35	0.65	hypothetical protein	SrbA	No hits
14003.77 [a]	T	G	NcR	0.37	0.79	---NA--	x	No hits
34174.84 [a,b]	C	T	CoR	0.36	0.66	nitrosoguanidine resistance	No hits	No hits
Low vs. Moderate								
18945.7	G	C	NcR	0.31	0.57	hypothetical protein	No hits	No hits
18945.9 [a,b]	C	T	NcR	0.38	0.92	hypothetical protein	No hits	No hits
14003.77 [b]	T	G	NcR	0.31	0.49	---NA--	x	No hits
46939.81 [a]	T	A	x	0.37	0.87	---NA--	x	GIT3
AUDPC								
18945_6	T	A	NcR	0.56	0.71	hypothetical protein	No hits	No hits
18945_8 [b]	A	T	NcR	-0.37	0.62	hypothetical protein	No hits	No hits
12430_32 [a,b]	A	G	NcR	-0.49	0.89	---NA--	x	No hits
1691_81	G	A	CoR	0.51	0.79	---NA---	No hits	No hits
28376_85 [a]	T	C	NcR	-0.54	0.73	hypothetical protein	x	No hits

a - SNPs also selected with a PIP $_{0.99;}$ b - SNPs identified as potentially associated in single and multi-association analyses; CR – coding Region; NCR – Non-coding region; x - no information

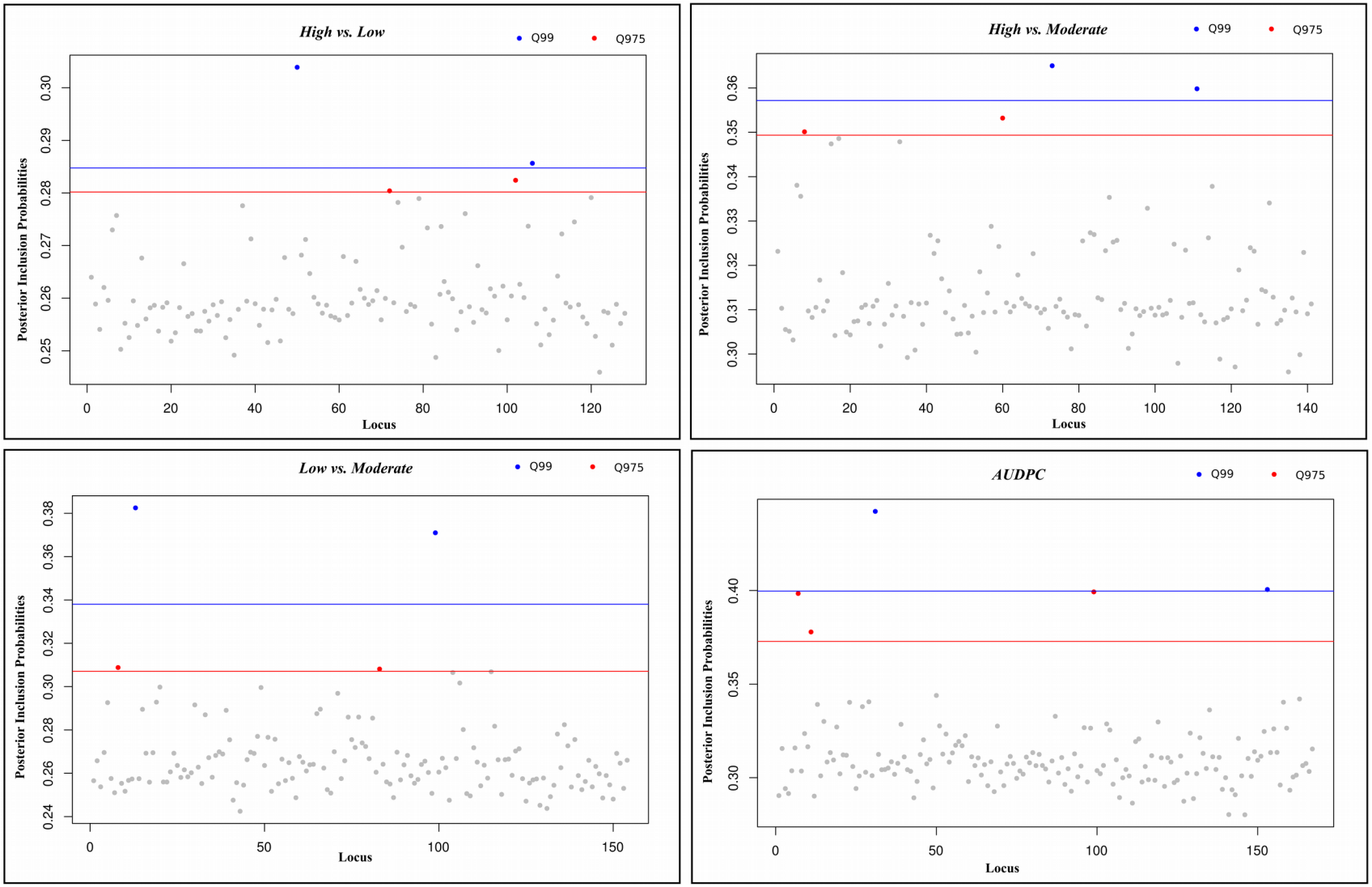


Figure 4.7 - Posterior inclusion probabilities (PIPs) for each SNP in each pairwise comparison in multi-SNP association test. The horizontal blue lines correspond to the PIP 99% empirical quantile threshold and red lines to the 97.5% empirical quantile. Blue dots: SNPs with a PIP > 99% empirical quantile, Red dots: SNPs with a PIP > 97.5% empirical quantile, Light grey dots: SNPs with a PIP < 97.5% empirical quantile

The search for the virulence role was performed in the PHI-base, in which the mutant phenotype of two genes was unable to affect the pathogenicity in other host-pathogen interactions, while two gene (GIT3, srbA) was shown to produce a mutant phenotype of reduced virulence in *Candida albicans* and *Aspergillus fumigatus*, respectively (**Table A3.5**). Finally, the majority of the SNPs putatively associated with aggressiveness into the single and multi association analysis are located in non-coding regions, being only 6 SNPs (34174.84; 44503.84; 17838.69; 1691.81; 41944.81; 35951.85) located in coding regions (**Table 4.1 and Table 4.3**)

4.5 Discussion

4.5.1 Phylogenetic relationships and host specialization

One of the first and most striking findings of this work was the high genetic differentiation, at the genomic level, between pathogenic and non-pathogenic fungi to green coffee berries (**Figure 4.3**), reinforcing the idea that *C. kahawae* should be considered as a distinct species. The phylogenetic analysis, besides confirming the clear pattern of population structure proposed in chapter 2, also revealed that *C. ciggaro* is the only paraphyletic group, and consequently, the possibility of this group being in fact more than one species can not be discarded and should be further investigated. The non-pathogenic isolate most closely related with *C. kahawae* was Cg126212 (ICMP18539) instead of Cg_432 as previously referred by Silva *et al.*, (2012), which reinforces the importance of using large number of loci to capture a more accurate phylogenetic relationship. The remaining phylogenetic results corroborate the taxonomic classification proposed by Weir *et al.*, (2012), and place *C. aotearoa* as the most distant species of the *C. gloeosporioides* complex under study.

4.5.2 Footprints of genomic adaptation and candidate genes for pathogenicity in *C. kahawae*

In this study we identified 5 560 diagnostic SNPs potentially involved in the pathogenicity of *C. kahawae* to green coffee berries. Although it is not probable that all these SNPs are related to this specific trait, the probability of finding the genetic

variation involved in *C. kahawae* 's pathogenicity is quite high. In fact, the enrichment in non-synonymous mutations found in this dataset (*filtered_dataset)* when compared to the *total_dataset* is pretty promissory, especially because this pattern was not observed for the total number of SNPs located in non-coding regions and/or coding regions. Functional annotation of these non-synonymous mutations (*ns_filtered_dataset*), as well as of the diagnostic SNPs (*filtered_dataset*) and *total_dataset,* revealed that for the 2nd level of the "biological process", "molecular function" and "cellular component" categories, only small differences were observed between the datasets. In "biological process" category, genes involved in "cellular component organization or biogenesis" are only present in *total_dataset*, while for the remaining datasets a small enrichment on the genes involved in "response to stimulus" was observed. In "molecular protein" category, the genes involved in "transporter activity" seems to be over represented in both filtered datasets (*filtered_dataset* and *ns_filtered_dataset*), which could suggest that transporters, such as "ABC transporters" and " Major Facilitator Superfamily", have an important role in the pathogenicity of *C. kahawae*. Additionally, a significant enrichment of specific GO terms was observed in the *ns_filtered_dataset* when compared to the *total_dataset*, particularly the "oxidation-reduction process" and "integral component of membrane". It is noteworthy that most of the genes associated with the term "integral component of membrane" are in fact transporters.

The analysis of dN/dS ratio on the *filtered_dataset* showed that 258 genes could be under positive selection (*ps_filtered_dataset*), from which 26 have more than one non-synonymous mutation. The potential role of these genes in fungal pathogenicity and virulence was assessed by a BLAST search against the PHI-base. A total of 30% (77 genes) had homology, and 15% (40 genes) were described as important for fungal pathogenicity and virulence when a mutant phenotype was produced. Five of them were identified as genes required for pathogenicity in other fungi, inducing mutant phenotypes of "total loss of pathogenicity" ("chitin synthase", "GTP-binding protein", "ABC transporter", "alpha-mannosyltransferase cmt1" and "cytochrome p450") and the remaining 36 genes, including the ones responsible for a change in virulence, are mainly involved in oxidative responses (for instance "cytochrome P450") and transport (mainly "ABC Superfamily" and "MFS transporters"). The importance of these two biological processes is well documented in the literature (Buiate *et al.*, 2017; Chen *et*

al., 2017; Liu *et al.*, 2017; Rao and Nandineni, 2017; Zeng *et al.*, 2018) and a comparative genomic analysis between two *Colletotrichum* species (*C. sublineola* and *C. graminicola*) in different hosts, showed an enrichment of proteins of these classes in the non conserved proteins dataset, with transporters being the most represented PFAM category (Buiate *et al.*, 2017). Overall, MFS transporters are the most common category of secondary carrier proteins. Members of this group are involved in the uptake of essential minerals and nutrients, also functioning as nutrient sensors, while others are responsible for the transport of various drugs and toxins (Liu *et al.*, 2017). Moreover, MFS transporters also play an important role in cellular resistance to oxidative stress in *Alternaria alternata* (Chen *et al.*, 2017) and in some cases can act as virulence factors (Liu *et al.*, 2017). In turn, ABC transporters confer tolerance by efflux of compounds across the membrane, thereby preventing an increase in intracellular concentration of toxic substances (Coleman *et al.*, 2011). The relative high abundance of these transporters in our dataset suggests that detoxification and/or production of toxic compounds in host-pathogen interaction may play an important role in the pathosystems *C. arabica - C. kahawae*, in accordance to Buiate *et al.*, (2017). Additionally, the high representation of cytochrome P450s proteins in our dataset reinforces this hypothesis, since these proteins have an important role in primary and secondary metabolism and fungal pathogenicity (Rao and Nandineni, 2017). In fact, the enrichment of genes related to oxidative responses can be a result of host-pathogen evolution, since reactive oxygen species (ROS) have been described as vital for stress responses, programmed cell death and plant defenses (Silva *et al.*, 2006). Finally, transcription factors (TFs) were also highly represented, especially “the fungal specific transcription factor” and “C6 transcription factor”, which could suggest that changes in gene expression patterns may be also important to *C. kahawae’s* pathogenicity.

In conclusion, the majority of genes found to be potentially under selection and with an enriched representation were associated with transporters, oxidative response, and signaling, suggesting an important role for these biological processes in the adaptation of *C. kahawae* to *C. arabica*, and providing candidate genes for evolutionary changes. Similar findings were also reported by Buiate *et al.*, (2017) and Rao and Nandineni, (2017) based on different genomic comparative analyses within the genus *Colletotrichum*, which suggest that these adaptive mechanisms could be associated to

varying aspects of each host environment, and to the secretion of or evasion to toxic secondary metabolites. In this sense, host specificity in closely related pathosystems of the *Colletotrichum* genus could be not only a matter of pathogen recognition, but also a much broader adaptation to the living host environment across the entire course of pathogen development, which has presumably occurred during co-evolution of the host and its pathogen.

Nevertheless, it is important to note that only around ~30% of the RAD loci contained in each dataset was mapped or annotated due to the lack of a properly annotated reference genome. Associated with this, all the genomic variation located in non-coding regions were not further studied, and consequently, high amount of information was not retrieved. Moreover, the reproductive characteristics and the emergence scenario of *C. kahawae*, make it difficult to separate the demographic signal from the selection pattern. In fact, it has been proposed that *C. kahawae* is a true clonal pathogen that has emerged by a host-jump from a non-pathogenic group (Silva *et al.*, 2012). In such scenario, *C. kahawae* was subjected to a strong disruptive selection during the first stages of the adaptation to *C. arabica*. According to Grünwald *et al.*, (2016), asexual reproduction could greatly amplify new advantageous mutations to extremely high frequencies along the entire genome by hitchhiking, and not just at the neighboring genes. This would eliminate polymorphisms and maintain only the intact genome of those individuals in the population having the favored mutations, evidencing a strong genetic bottleneck and the lack of shared polymorphisms with the remaining non-pathogenic fungi. Thus, in a perfectly clonal pathogen, each adaptive allele that arises, will be linked to every other allele in the genome, and consequently the selection is more likely to act at the level of individual clones than individual alleles (Grünwald *et al.*, 2016; Plissonneau *et al.*, 2017; Shapiro *et al.*, 2009). In this sense, if the goal is to distinguish adaptive loci from other fixed mutations in the clonal background, the typical genome-scan may be a limited approach, and hence it is crucial to look for the excess of functional changes, such as the enrichment of non-synonymous mutations and search for genes putatively under selection (Plissonneau *et al.*, 2017; Shapiro *et al.*, 2009).

4.5.3 Genome-wide association of aggressiveness in *C. kahawae*

Based on our previous phenotypic evaluation of 26 *C. kahawae* isolates (Chapter 3), a genome-wide association analysis was performed in order to better understand the genetic mechanisms underlying aggressiveness. According to (Dalman *et al.*, 2013), applying a GWAS to an organism in a haploid stage, which can be clonally reproduced in high numbers and phenotyped repeatedly, increases the accuracy of the phenotypic measurements and the power of the association analysis in several orders of magnitude, when compared to diploids. For this reason, a low number of haploid individuals can be used to successfully identify robust associations in small fungal genomes, whereas several hundreds of individuals are needed in diploid organisms with large genomes such as humans and plants (Dalman *et al.*, 2013).

In *C. kahawae,* the number of SNPs that are not associated with population structure is low (173 SNPs) and correspond to only 5% of the total genetic variation. No causal SNPs were identified in this dataset, but instead a group of SNPs of small effect was detected. In single-SNP association analyses, the 10 individual SNPs found to be associated (BF0.97) showed a low phenotypic effect in aggressiveness (β| = 0.059 to |β| = 0.087) for all the pairwise analyses (*High vs. Moderate, High vs. Low, Low vs. Moderate)*, but in the continuous analysis (*AUDPC*) 5 SNPs present a moderate phenotypic effect (β| = 1.24). Despite these differences, three of the SNPs associated with aggressiveness in the continuous analysis were also found in the pairwise analyses.

In the multi-SNP association test, 15 SNPs with a posterior inclusion probability of 97.5% (PIP0.975) were found, showing moderate effect for all pairwise analyses (*High vs. Moderate, High vs. Low, Low vs. Moderate)* [PIP0.97 SNPs | β | > 0,36] and for the continuous analysis (*AUDPC*) [PIP0.97 SNPs | β | > 0,62]. The phenotypic effect seems to be smaller in all pairwise analyses that included the *High* aggressiveness class, probably due to the low number of isolates that compose this class. Moreover, only 7 SNPs were common to the SNPs identified in single SNP analyses, specifically 1 in 5 SNPs was common to both approaches in *High vs. Low* and *High vs. Moderate* pairwise analyses, 3 in 4 for *Low vs. Moderate* pairwise analysis, and 3 in 5 for the continuous

analysis with *AUDPC*. These results provide additional support for the reliability of the *Low vs. Moderate* and *AUDPC* analyses.

The annotation of the loci containing the SNPs putatively associated with aggressiveness allowed the identification of three candidate genes in the single-association analysis ("F-box domain-containing", "nitrosoguanidine resistance" and "Fungal specific transcription factor domain-containing"), and two candidate genes in the multi-SNP association analysis ("nitrosoguanidine resistance" and "C6 transcription factor"). The "nitrosoguanidine resistance", which encodes for an integral component of the membrane that is able to regulate the fungal-type cell wall organization and phospholipid translocation (García-López *et al.*, 2010; García-Marqués *et al.*, 2016), was the only candidate gene common to both approaches. In addition, among the few candidate genes identified, two are transcription factors located in the nucleus, which may suggest that differential gene expression and/or associated regulatory mechanisms might have a preponderant role in aggressiveness. However, none of these genes was previously associated with aggressiveness in other plant-pathogen interactions, which suggests that they may be specific of *C. arabica – C. kahawae* interaction. The remaining significantly associated SNPs were hypothetical proteins, or are located in intergenic regions, or not annotated. These SNPs may represent regulatory elements or unknown genes that are responsible for the trait variance.

Finally, the success and power of an association study relies on the number of SNP markers and on the LD decay. In *C. kahawae,* both conditions are far from ideal, as a low number of SNPs not related with the population structure pattern was identified and the entire genome is inherited asexually as a single, non-recombining linkage group, which can increase the number of false positives. Therefore, as in any GWA study, a further detailed investigation of these candidate genes is required and will allow to confirm their involvement in *C. kahawae's* aggressiveness and assess their causative effect at the phenotypic level.

4.6 Conclusions

This work took the first step towards the understanding of the genetic mechanisms underlying the ability of *C. kahawae* to infect green coffee berries and its

aggressiveness. Our results suggest that *C. kahawae's* pathogenicity involves several biological processes such as, detoxification and transport, regulation of host and pathogen gene expression, and signaling. 15% of the genes potentially under selection were described as having an important role in fungal pathogenicity and virulence, being some of them identified as genes responsible for the total loss of pathogenicity in other fungi. Finally, the high abundance of TF may suggest that expression changes in gene expression patterns can be more important than the presence/absence of individual gene alleles. On the other hand, aggressiveness does not seem to be regulated by any causal mutation and even the associated SNPs are of small effect, which leads to three possible conclusions: **i)** aggressiveness is regulated by a set of many small effect SNPs difficult to detect with a GWAS analysis; **ii)** aggressiveness is a plastic trait regulated by epigenetic features (differential gene expression and associated regulatory mechanisms), consequently, a transcript analysis is needed to complement the current study; **iii)** aggressiveness is not under selection and is governed by physiological conditions. Nevertheless, a repertoire of candidate genes is now provided and can be studied through gene expression and additional functional analyses (knockouts, knockdowns and transgenics) to ascertain their causative role in *C. kahawae* aggressiveness and pathogenicity. Finally, the collected information could used in the development of future evidence-based sustainable control measures.

4.7 References

Batista, D., Silva, D.N., Vieira, A., *et al*. (2017) Legitimacy and implications of reducing *Colletotrichum kahawae* to subspecies in plant pathology. *Front. Plant Sci.* **7**, 1–4.

Bridge, P.D., Waller, J.M., Davies, D. and Buddie, A.G. (2008) Variability of *Colletotrichum kahawae* in relation to other *Colletotrichum* species from tropical perennial crops and the development of diagnostic techniques. *J. Phytopathol.* **156**, 274–280.

Buiate, E.A.S., Xavier, K. V, Moore, N., Torres, M.F., Farman, M.L., Schardl, C.L. and Vaillancourt, L.J. (2017) A comparative genomic analysis of putative pathogenicity genes in the host-specific sibling species *Colletotrichum graminicola* and *Colletotrichum sublineola*. *BMC Genomics* **18**, 67.

Byers, K., Xu, S. and Schlüter, P.M. (2016) Molecular mechanisms of adaptation and speciation: why do we need an integrative approach? *Mol. Ecol.* **26**, 277-290.

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A. and Cresko, W. a. (2013) Stacks: An analysis tool set for population genomics. *Mol. Ecol.* **22**, 3124–3140.

Chen, L., Tsai, H., Yu, P. and Chung, K. (2017) A Major Facilitator Superfamily transporter-mediated resistance to Oxidative Stress and fungicides requires Yap1, Skn7, and MAP Kinases in the Citrus fungal pathogen *Alternaria alternata*. *PLoS One* **12**, e0169103.

Coleman, J.J., White, G.J., Rodriguez-carres, M. and Vanetten, H.D. (2011) An ABC transporter and a Cytochrome P450 of Nectria haematococca MPVI are virulence factors on pea and are the major tolerance mechanisms to the phytoalexin pisatin. *Mol Plant Microbe Interact.* **24**, 368–376.

Connelly, C.F. and Akey, J.M. (2012) On the prospects of Whole-Genome association mapping in *Saccharomyces cerevisiae*. *Genetics* **191**, 1345–1353.

Cooke, D.E.L., Cano, L.M., Raffaele, S., *et al*. (2012) Genome analyses of an aggressive and invasive lineage of the Irish Potato Famine pathogen. *PLoS Pathog.* **8**, e1002940.

Dalman, K., Himmelstrand, K., Olson, Å., Lind, M., Brandström-Durling, M. and Stenlid, J. (2013) A Genome-Wide association study identifies genomic regions for virulence in the non-model organism *Heterobasidion annosum* s.s. *PLoS One* **8**, e53525.

Danecek, P., Auton, A., Abecasis, G., *et al*. (2011) The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158.

Delmas, E.L.C., Fabre, F., Jérôme, J., Mazet, I.D., Cervera, S.R., Laurent, D. and François, D. (2016) Adaptation of a plant pathogen to partial host resistance: selection for greater aggressiveness in grapevine downy mildew. *Evol. Appl.* **9**, 709–725.

Eaton, D.A.R. (2014) PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics* **30**, 1844–1849.

Elad, Y. and Pertot, I. (2014) Climate change impacts on plant pathogens and plant diseases. *J. Crop Improv.* **28**, 99–139.

Excoffier, L. and Lischer, H.E.L. (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* **10**, 564–567.

Gotz, S., García-Gómez, J.M., Terol, J., Williams, T.D., Go, S., Talo, M., Nagaraj, S.H. and Conesa, A. (2008) High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* **36**, 3420–3435.

Grünwald, N.J., McDonald, B.A.M. and Milgroom, M.G.M.G. (2016) Population genomics of fungal and Oomycete Pathogens. *Annu. Rev. Phytopathol.* **54**, 323–346.

Guan, Y. and Stephens, M. (2011) Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Ann. Appl. Stat.* **5**, 1780–1815.

Jonge, R. de, Esse, H.P. van, Maruthachalam, K., *et al*. (2012) Tomato immune receptor Ve1 recognizes effector of multiple fungal pathogens uncovered by genome and RNA sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 5110–5.

Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079.

Liu, L., Yan, Y., Huang, J., Hsiang, T., Wei, Y., Li, Y., Gao, J. and Zheng, L. (2017) A Novel MFS transporter gene ChMfs1 is important for hyphal morphology, conidiation, and pathogenicity in *Colletotrichum higginsianum*. *Front. Microbiol.* **8**, 1–11.

Loureiro, A., Guerra-Guimarães, L., Lidon, F.C., Bertrand, B., Silva, M.C. and Várzea, V. (2011) Isoenzymatic characterization of *Colletotrichum kahawae* isolates with different levels of aggressiveness. *Trop. Plant Pathol.* **36**, 287–293.

Miller, M.A., Pfeiffer, W. and Schwartz, T. (2010) Creating the CIPRES Science Gateway for inference of large phylogenetic trees. *2010 Gatew. Comput. Environ. Work. GCE 2010*.

Möller, M. and Stukenbrock, E.H. (2017) Evolution and genome architecture in fungal plant pathogens. *Nat. Rev. Microbiol.* **15**, 756–771.

Pariaud, B., Ravigné, V., Halkett, F., Goyeau, H., Carlier, J. and Lannou, C. (2009) Aggressiveness and its role in the adaptation of plant pathogens. *Plant Pathol.* **58**, 409–424.

Pires, A.S., Azinheira, H.G., Cabral, A., *et al*. (2016) Cytogenomic characterization of *Colletotrichum kahawae*, the causal agent of coffee berry disease, reveals diversity in minichromosome profiles and genome size expansion. *Plant Pathol.* **65**, 968–977.

Plissonneau, C., Benevenuto, J., Mohd-Assaad, N., Fouché, S., Hartman, F.E. and Croll, D. (2017) Using population and comparative genomics to understand the genetic basis of Effector-Driven fungal pathogen evolution. *Front. Plant Sci.* **8**, 119.

Posada, D. and Buckley, T.R. (2004) Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Syst. Biol.* **53**, 793–808.

Rao, S. and Nandineni, M.R. (2017) Genome sequencing and comparative genomics reveal a repertoire of putative pathogenicity genes in chilli anthracnose fungus *Colletotrichum truncatum*. *PLoS One* **12**, e0183567.

Ronquist, F., Teslenko, M., Mark, P. Van Der, *et al*. (2012) Mrbayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539–542.

Schmidt, S.M., Lukasiewicz, J., Farrer, R., Dam, P. van, Bertoldo, C. and Rep, M. (2016) Comparative genomics of *Fusarium oxysporum* f. sp. *melonis* reveals the secreted protein recognized by the Fom-2 resistance gene in melon. *New Phytol.* **209**, 307–318.

Servin, B. and Stephens, M. (2007) Imputation-based analysis of association studies: Candidate regions and quantitative traits. *PLoS Genet.* **3**, 1296–1308.

Shapiro, B.J., David, L.A., Friedman, J. and Alm, E.J. (2009) Looking for Darwin's footprints in the microbial world. *Trends Microbiol.* **17**, 196–204.

Silva, C., Várzea, V., Guerra-guimarães, L., Azinheira, H.G., Fernandez, D., Petitot, A., Bertrand, B., Lashermes, P. and Nicole, M. (2006) Coffee resistance to the main diseases : leaf rust and coffee berry disease. *Braz. J. Plant Physiol.* **18**, 119–147.

Silva, D.N., Talhinhas, P., Cai, L., Manuel, L., Gichuru, E.K., Loureiro, A., Várzea, V., Paulo, O.S. and Batista, D. (2012) Host-jump drives rapid and recent ecological speciation of the emergent fungal pathogen *Colletotrichum kahawae. Mol. Ecol.* **21**, 2655–2670.

Stamatakis, A. (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313.

Talas, F., Kalih, R., Miedaner, T. and McDonald, B.A. (2016) Genome-Wide association study Identifies novel candidate genes for aggressiveness, Deoxynivalenol production, and Azole sensitivity in natural field populations of *Fusarium graminearum*. *Mol. Plant. Microbe. Interact.* **29**, MPMI-09-15-0218.

Urban, M., Cuzick, A., Rutherford, K., *et al*. (2017) PHI-base: a new interface and further additions for the multi-species pathogen–host interactions database. *Nucleic Acids Res.* **45**, D604–D610.

Várzea, V.M.P, Rodrigues, J.C. and Lewis, B. (2002) Distinguishing characteristics and vegetative compatibility of *Colletotrichum kahawae* in comparison with other related species from coffee. *Plant Pathol.* **51**, 202–207.

Weir, B.S., Johnston, P.R. and Damm, U. (2012) The *Colletotrichum gloeosporioides* species complex. *Stud. Mycol.* **73**, 115–180.

Zeng, Z., Sun, H., Vainio, E.J., Raffaello, T., Kovalchuk, A., Morin, E., Duplessis, S. and Asiegbu, F.O. (2018) Intraspecific comparative genomics of isolates of the Norway spruce pathogen (*Heterobasidion parviporum*) and identification of its potential virulence factors. *BMC Genomics* **19**, 220

Zhan, J., Thrall, P.H. and Burdon, J.J. (2014) Achieving sustainable plant disease management through evolutionary principles. *Trends Plant Sci.* **19**, 570–575.

Zhan, J., Thrall, P.H., Papaïx, J., Xie, L. and Burdon, J.J. (2015) Playing on a pathogen's Weakness: Using evolution to guide sustainable plant disease control strategies. *Annu. Rev. Phytopathol.* **53**, 2.1-2.25.

Comparative validation of conventional and RNA-seq data-derived reference genes for qPCR expression studies of *Colletotrichum kahawae*

5

Vieira A.[a,b,c], Cabral A.[c], Fino J.[b], Azinheira HG.[a,c], Loureiro A.[a,c], Talhinhas P.[a,c,d], Pires AS.[a,d], Várzea V.[a,c], Moncada P.[d], Oliveira H.[c], Silva MC.[a,c], Paulo OS.[b], Batista D.[a,b,c]

[a]**CIFC/ISA** - UL, Oeiras, Portugal; [b]**CoBiG2/cE3c/FCUL** - UL, Lisboa, Portugal; [c]**LEAF/ISA -** UL, Lisboa, Portugal, [d]**PCB/ITQB -** UNL, Oeiras, Portugal, [e]**Cenicafe –** Manizales, Colombia

5.1 Abstract

Colletotrichum kahawae is an emergent fungal pathogen causing severe epidemics of Coffee Berry Disease on Arabica coffee crops in Africa. Currently, the molecular mechanisms underlying the *Coffea arabica* - *C. kahawae* interaction are still poorly understood, as well as the differences in pathogen aggressiveness, which makes the development of functional studies for this pathosystem a crucial step. Quantitative real time PCR (qPCR) has been one of the most promising approaches to perform gene expression analyses. However, proper data normalization with suitable reference genes is an absolute requirement. In this study, a set of 8 candidate reference genes were selected based on two different approaches (literature and Illumina RNA-seq datasets) to assess the best normalization factor for qPCR expression analysis of *C. kahawae* samples. The gene expression stability of candidate reference genes was evaluated for four isolates of *C. kahawae* bearing different aggressiveness patterns (Ang29, Ang67, Zim12 and Que2), at different stages of fungal development and key time points of the plant-fungus interaction process. Gene expression stability was assessed using the pairwise method incorporated in geNorm and the model-based method used by NormFinder software. For *C. arabica* - *C. kahawae* interaction samples, the best normalization factor included the combination of *PP1*, *Act* and *ck34620* genes, while for *C. kahawae* samples the combination of *PP1*, *Act* and *ck20430* revealed to be the most appropriate choice. These results suggest that RNA-seq analyses can provide alternative sources of reference genes in addition to classical reference genes. The analysis of expression profiles of bifunctional catalase-peroxidase (*cat2*) and trihydroxynaphthalene reductase (*thr1*) genes further enabled the validation of the selected reference genes. This study provides, for the first time, the tools required to conduct

accurate qPCR studies in *C. kahawae* considering its aggressiveness pattern, developmental stage and host interaction.

5.2 Introduction

Colletotrichum kahawae J.M. Waller & P.D. Bridge, the causal agent of Coffee Berry Disease (CBD), is a highly aggressive and specialized fungal pathogen of coffee. This pathogen currently occurs in nearly all African regions where Arabica coffee (Coffea Arabica L.) is grown, particularly at high altitudes, ravaging plantations and causing up to 80% yield losses, if no control measures are applied (Gichuru *et al.*, 2008; Hindorf and Omondi, 2011; van der Vossen *et al.*, 1976). The potential introduction of this quarantine pathogen into other continents represent a major concern and threat, particularly to high altitude Arabica coffee plantations that can also be found in Latin America and Asia. *C. kahawae* infects several coffee organs but maximum production losses occur when infection takes place in expanding green berries, leading to their premature dropping and mummification (Hindorf and Omondi, 2011; Silva *et al.*, 2006; van der Vossen *et al.*, 1976).

Functional and molecular studies to better understand *C. arabica* - *C. kahawae* interactions are of key importance to aid disease resistance breeding efforts. Real time quantitative PCR (qPCR) is currently the most accurate method for analyzing gene differential expression given its capability of detecting low abundance mRNAs and slight differences in the expression level. When compared to other methods used to evaluate transcript accumulation, such as Northern blotting, RNase protection assay, *in situ* hybridisation, and cDNA microarray technology, the main advantages of qPCR are its high sensitivity and specificity. However, in order to obtain reliable and reproducible results, proper data normalization is necessary (Bustin *et al.*, 2010). Different normalization strategies have been proposed (Huggett *et al.*, 2005; Wong and Medrano, 2005), but normalization of gene expression levels by reference genes (RGs) is most certainly the "gold standard", though the success of this procedure relies on the appropriate choice of RGs (Andersen *et al.*, 2004; Vandesompele *et al.*, 2002). Typically, RGs consist of a group of constitutively expressed genes which are considered to be essential to maintain basic cellular functions, and are ubiquitously expressed in all cells of an organism, irrespective of tissue type, developmental stage, cell cycle state, or external signals (Lin *et al.*, 2014). However, some evidence shows

that almost all genes seem to be regulated at some point under certain conditions and there are always some variations in transcript levels, so that none of the commonly exploited genes can be viewed as universal RGs (Andersen *et al.*, 2004; Vandesompele *et al.*, 2002). So far, most studies have focused on validating a subset of commonly used RGs for a specific context. Although this seems to be a good strategy, such studies try to identify the best candidates from a small and "a priori" set of RGs, assuming that at least one or a few of them are suitable for the experimental context under study (Hruz *et al.*, 2011). In order to avoid this potential bias, researchers have conducted genome-wide surveys to search for new RGs using different types of datasets such as microarray (Hruz *et al.*, 2011), EST libraries (Coker and Davies, 2003; Zhu *et al.*, 2008), tag-based approach serial analysis of gene expression (SAGE) (Velculescu *et al.*, 1999) and more recently RNA-seq (Kim and Yun, 2011; Lin *et al.*, 2014; Llanos *et al.*, 2015). RNA-seq seems to be a powerful tool for identifying reference genes across a global transcriptome, providing a significantly more accurate measurement of transcript abundance (Lin *et al.*, 2014). Recently, a high-throughput RNA sequencing (RNA-seq) approach was applied to study compatible and incompatible *Coffea* spp. - *C. kahawae* interactions providing unprecedented information on the coding-genes putatively involved, and on their expression for both the host and the pathogen (Fino *et al.*, 2014). RGs suitable for gene expression studies in both resistant and susceptible coffee genotypes to *C. kahawae* were already established (Figueiredo *et al.*, 2013), but no normalization system was developed to study gene expression in the pathogen. Moreover, since the validation of RNA-seq data is ideally achieved through qPCR analysis, the establishment of a set of RGs for the pathogen is a crucial step to validate the fungal RNA-seq results.

In this study, we selected a set of 8 candidate RGs, including RGs previously used for other phytopathogenic fungi and a new set of genes retrieved from the analysis of several RNA-seq datasets of a compatible and incompatible *Coffea* spp. - *C. kahawae* interaction (Fino *et al.*, 2014). The stability of the candidate RGs was evaluated using four *C. kahawae* isolates bearing different aggressiveness patterns in a range of *C. kahawae* samples and *C. arabica* - *C. kahawae* samples. Furthermore, since *C. kahawae* is a hemibiotrophic pathogen with substantial biomass variation during the infection stages within the plant (Loureiro *et al.*, 2012), an additional methodology for

normalization of the *C. kahawae* biomass in *C. arabica* - *C. kahawae* samples was required. For that, the biomass quantification was carried out by measuring the fungal DNA by qPCR in the same *C. arabica* - *C. kahawae* samples used and accounted for in the expression analyses.

The best combination of RGs determined for each dataset was further validated by assessing the expression of two genes described as putatively involved in the early steps of the infection process of pathogenic fungi, namely a bifunctional catalase-peroxidase (*cat2*) and a trihydroxynaphthalene reductase (*thr1*) (Bourdais *et al.*, 2012; Brown *et al.*, 2008; Perpetua *et al.*, 1996; Tanabe *et al.*, 2011; Thompson *et al.*, 2000; Tsuji *et al.*, 2003; Yarden *et al.*, 2014). *cat2* was described as being involved in reactive oxygen species (ROS) detoxification by removing intracellular hydrogen peroxide (H2O2) and converting it into water and dioxygen (Bourdais *et al.*, 2012; Brown *et al.*, 2008; Tanabe *et al.*, 2011; Yarden *et al.*, 2014). This protein has a crucial role in the first stages of the infection process, since previous studies revealed its involvement on pathogen protection against the generation of ROS by the host, one of the most rapid and dramatic defense reactions activated by the plant following the pathogen attack (Tanabe *et al.*, 2011). The trihydroxynaphthalene reductase protein encodes the *thr1* gene and was previously described as being involved in the first reduction step of the melanin biosynthesis pathway. Melanin is important for pathogenicity in some plant pathogenic fungi since appressorial melanization is essential for penetration of host tissues (Bourdais *et al.*, 2012; Perpetua *et al.*, 1996; Thompson *et al.*, 2000; Tsuji *et al.*, 2003). The expression profiles evaluation of those genes with a different set of candidate RGs clearly demonstrates the impact of different normalization approaches in the final results. The global strategy applied on this work allowed, for the first time, the establishment of a new set of RGs suitable for gene expression studies in *C. kahawae*, considering its aggressiveness pattern or stages of fungal development and infection.

5.3 Material and Methods

5.3.1 Fungal isolates

Four *C. kahawae* isolates (CIFC/ISA/Universidade de Lisboa collection) bearing different aggressiveness patterns, as observed in CIFC´s routine screening tests, were

used (**Table 5.1**): Ang 29 and Zim 12- highly aggressive isolates (leading to tissue complete necrosis in 6-8 dpi); Que 2- moderately aggressive isolate (leading to tissue complete necrosis in 10 dpi); and Ang 67- low aggressive isolate (leading to tissue complete necrosis in 20 dpi). These isolates were grown in 90 mm polystyrene Petri dishes containing malt extract agar (MEA, Oxoid, England) for 7 days under a photoperiod of 12 h at 22°C, in order to obtain conidia.

Table 5.1 - Details on *C. kahawae* isolates regarding its geographical origin and aggressiveness pattern in *Coffea arabica* (var. Caturra).

Isolate	Origin	Aggressiveness pattern
Ang29	Angola	High
Ang67	Angola	Low
Que2	Kenya	Medium
Zim12	Zimbabwe	High

5.3.2 Inoculation of coffee hypocotyls

Coffee hypocotyls were used as a model material in controlled conditions, since previous studies have shown a correlation between the pre-selection test on hypocotyls and mature plant resistance to CBD in the field (r =0.73–0.80) (van der Vossen *et al.*, 1976). Seeds of *C. arabica* (var. Caturra; susceptible to CBD) were sown in seedbeds in a growth chamber (FITOCLIMA Walk-in 10000 EHHF, Aralab, Portugal) under controlled conditions (24-26°C, with 12 h photoperiod at 800 $\mu mol.m^{-2}.s^{-1}$ light and 75-85%relative humidity) during 7-8 weeks. Plantlets were collected after emergence (prior to cotyledon expansion) and hypocotyls were then inoculated and maintained as described by Figueiredo *et al.* (2013). Briefly, hypocotyls were sprayed with a conidial suspension of *C. kahawae* (3×106 $conidia.ml^{-1}$) and maintained in a moist chamber at 22°C in the dark for 24 h, and then under a 12 h photoperiod during the infection time-course.

5.3.3 Sample preparation and collection

In this work, two types of samples were collected: *C. kahawae* samples, representing different stages of fungal development [ungerminated conidia, germinated conidia with mature appressoria formed after 18-22h on polystyrene or on leaves of *C. arabica*

(hereafter referred as appressoria), and saprophytic mycelium] and *C. arabica* - *C. kahawae* samples, representing key steps of the infection process (melanized appressoria, beginning of penetration and early stages of necrotrophy).

5.3.3.1 *C. kahawae* samples:

Ungerminated conidia were harvested from 7-to 10-day-old cultures, grown on MEA medium under a photoperiod of 12 h, into sterile water and cleared from mycelium using a borosilicate glass filter crucible with porosity 1. To produce saprophytic mycelium, 5 ml of conidial suspension (2 × 106 conidia.ml^{-1}) were inoculated in 25 ml of Potato Dextrose Broth (BD-Difco, USA) and grown for 3 days at 22°C. The mycelium was filtered off from the culture broth through a fine cloth mesh and washed with distilled water. The mycelium was lyophilized and stored at -80°C until use. *In vitro* appressoria were obtained following the methodology of (Kleemann *et al.*, 2008) with some modifications. A monolayer of 15 ml conidial suspension (1 × 106 conidia ml^{-1}) was used and filter paper was applied to the liquid surface instead of the nylon mesh. The appressoria formed on polystyrene plates after 18-22 h were scrapped with 10 ml of sterile water containing 0.02% Tween 20 (Fisher Scientific, USA). The suspension was centrifuged at 5000 g for 10 min, and the appressoria pellet was collected, lyophilized and stored at -80°C. To obtain appressoria *in planta*, the abaxial surface of 12 young leaves of *C. arabica* (var. caturra) were sprayed with a conidial suspension (2 × 106 conidia.ml^{-1}) of each isolate, using an atomizer. The inoculated leaves were maintained in a humidity box for 24h at 22°C to allow conidia germination and appressoria formation. After this period, the leaves were air dried and a thin layer of nail polish was applied (Loureiro *et al.*, 2015). The nail polish, containing the fungal structures [as verified by microscopic examination of cotton blue-stained nail polish fragments (Silva *et al.*, 1999)] was allowed to dry for 24h and then carefully removed from the leaves, and stored at -80°C.

5.3.3.2 *C. arabica* - *C. kahawae* samples:

Hypocotyls of *C. arabica* (var. caturra) were inoculated with a conidial suspension of each *C. kahawae* isolate under study, as described above. To determine sampling times corresponding to the key time points of pathogenesis for the different *C. kahawae*

isolates, the fungal pre-penetration, penetration and post-penetration stages were evaluated by light microscopy, as previously described (Loureiro *et al.*, 2012; Silva *et al.*, 1999). Plant material was harvested accordingly with the stages referred above: **i)** differentiation of melanized appressoria at 24 hours post inoculation (hpi) for all *C. kahawae* isolates; **ii)** fungal penetration and establishment of the biotrophic phase at 48 hpi for Ang29, Que2 and Zim12 but only at 72 hpi for Ang67; **iii)** switch to necrotrophy (onset of first symptoms) at 72 hpi for all isolates excluding Ang67, for which the first lesions only appeared at 96 hpi. Symptoms were recorded during the entire infection process, until the death of all hypocotyls, to confirm the aggressiveness profile of the isolates used. Two independent experiments were conducted and 40 hypocotyls were collected for each isolate-time point combination. Hypocotyls were immediately frozen by immersion in liquid nitrogen and stored at -80°C.

5.3.4 RNA extraction and cDNA synthesis

Total RNA was extracted with Spectrum™ Plant Total RNA Kit (Sigma-Aldrich, USA) according to the manufacturer's instructions, for all samples. Residual genomic DNA was digested with DNase I (On-Column DNase I Digestion Set, Sigma-Aldrich, USA). Lyophilized samples of conidia and appressoria formed in vitro were grounded in the Lysis Solution with sand, while samples containing mycelium, nail polish with appressoria formed in vivo and infected hypocotyls were grounded in liquid nitrogen. RNA purity and concentration were measured at 260/280 nm and 260/230 nm using a spectrophotometer (NanoDrop-1000, Thermo Scientific, USA), while RNA integrity was verified by agarose gel electrophoresis. Genomic DNA contamination on the crude RNA samples was verified by qPCR analysis in an iQ5 real-time thermalcycler (Bio-Rad, USA), using EvaGreen® Supermix (Bio-Rad). Each 15 µl reaction comprised 5 µl of crude RNA, 7.5 µl EvaGreen Supermix and 200 nM of each ck39066 primer. First-strand cDNA was synthesized from 1.0 µg of total RNA for *C. kahawae* samples and from 1.7 µg of total RNA for *C. arabica* - *C. kahawae* samples in a 20 µl final volume, using Omniscript RT kit (Qiagen, Germany) and Oligo(dT)18 primer (MBI Fermentas, Lithuania), following the manufacturer's instructions. The cDNA was diluted 1:25 with sterile water for *C. kahawae* samples, and 1:20 for *C. arabica* - *C. kahawae* samples and stored at -20°C.

5.3.5 Selection of candidate reference genes

The candidate RGs were selected based on two different approaches: three were selected from the literature as the most promising RGs for related plant pathogenic fungi; and five were selected from Illumina RNA-seq datasets, representing the early events of a compatible and an incompatible *Coffea* sp.- *C. kahawae* interaction (24, 48 and 72 hpi), based on gene expression stability (Fino *et al.*, 2014). The RGs retrieved from the literature were selected considering different functional classes, in order to reduce the chance of co-regulation and thus avoid a significant bias in geNorm analysis. These included: actin (*Act*) (Fang and Bidochka, 2006; Zhou *et al.*, 2012); cyclophilin type peptidyl-prolyl cis-trans isomerase precursor (*Cyp*) (Kim and Yun, 2011); and serine threonine-protein phosphatase (*PP1*) (Zhou *et al.*, 2012). For these genes it was necessary to obtain the DNA sequence for *C. kahawae* in order to design species specific qPCR primers. Novel sequences were lodged in GenBank with accession numbers KU579251 to KU579253. For the Illumina RNA-seq-based approach, candidate RGs were selected among the 653 *C. kahawae* genes previously identified (Fino *et al.*, 2014). The candidate RGs were ranked according to the following criteria: presenting a fold change between sample collection times (24, 48, 72 hpi) close to one and low standard errors (tested pairs: 24/48 hpi, 48/72 hpi and 24/72 hpi for both compatible and incompatible interactions), i.e. with a constitutively expression along the infection period studied. The top six ranked genes selected under these criteria were *ck20430* (predicted 60S ribosomal protein L18 gene); *ck28444* (predicted membrane biogenesis protein yop1); *ck36020* (predicted stf2-like protein); *ck48742* (predicted 40S ribosomal protein S28); and *ck34620* (homologue to hypothetical protein CGGC5_3535).

Specific primers (**Table 5.2**) were designed for each gene with PerlPrimer v1.1.17 (Marshall, 2004). Whenever possible, primers were designed in the junction of two different exons, thus preventing amplification of potential residual DNA. Amplicon size ranged between 80-200bp for all genes used, except for *PP1*, *cat2* and *thr1* for which longer fragments had to be produced (222- 269pb) in order to ensure the design of good quality primers (**Table 5.2**).

Table 5.2 - Detailed description of candidate reference genes, genes of interest, primer sets and qPCR amplification conditions

Name	Description	Go terms Molecular Function	Primer sequence (5'-3')	Amplicon length (bp)	Annealing temperature (ºC)	PCR efficiency	Primers designed in intron-exon boundary?
Act	Actin	GO:0005524; GO:0005200	F - CAACATTGTCATGTCTGGTGG R - GTACTCCTGCTTGGAGATCC	202	63	1,902	YES, F
Cyp	Cyclophilin type peptidyl-prolyl cis-trans isomerase B precursor	GO:0003755	F - AAGACCGCTGAGAACTTCCG R - CTCGCCGTAGATGGACTTGC	153	64	1,910	YES, R
PP1	Serine threonine-protein phosphatase	GO:0004722	F- CACTGGTTGGAGCGAAAACG R - CAGGATCTGGAACGAGCAAAG	250	60	1,919	YES, R
ck20430	60s ribosomal protein L18	GO:0003735	F- AGAGACCAACAGCACCACAC R-CCACAAGCACAAGAAGCCC	118	60	1.923	YES, R
ck28444	membrane biogenesis protein yop1	GO:0034613; GO:0048309; GO:0071786; GO:1902408; GO:0051292; GO:0016192	F- TAACAACCTCGAGAAGCAGACC R- CGACCCAGTAAGTCAGCCAC	206	60	1.935	YES, R
ck48742	40S ribosomal protein S28	GO:0003735	F-ACCAGACCCGTTCCATCATCC R- CAAGATCCGTTCCTCGTAATGTCC	158	58	1.929	YES, F
ck36020	stf2-like protein	None	F-CCACGGCCCCAACGAGGAGGAT R- GAGGGCTGCAGCACGAAACATTAGG	140	60	1.930	NO
ck39066	homologue to hypothetical protein CGGC5_4189	None	F-AAGGGTGAATGGTTGAAGGG R- CTGCGTATGGGAAGAAGTAGAC	151	60	1.907	NO
ck34620	homologue to hypothetical protein CGGC5_3535	None	F-CCCGACTTCCACTTCCATTACC R-CGCCGACCAGGATGAACTTG	208	63	1.926	NO
ck21238	bifunctional catalase-peroxidase cat2	GO:0004096; GO:0004601; GO:0016491; GO:0020037; GO:0046872	F- TTCCGCATCTACCTTCCG R- TCAACACCAGCAACACCAC	222	60	1.950	NO
ck25805	trihydroxynaphthalene reductase	GO:0008152; GO:0055114	F- ATGTACCGTGATGTCTGCC R- GTGATCCAATCTACTAATACCAGCC	269	60	1.921	YES, F

The sequence of cDNA was obtained from the Illumina RNA-seq database (Fino *et al.*, 2014) and the DNA sequence was inferred based on the alignment with sequences from other *Colletotrichum* species, and subsequently verified upon sequencing of the amplicon obtained for *C. kahawae* (**Table A4.1**). The specificity of real-time PCR products was confirmed by the presence of a single peak in the melting curve and the presence of a single band with the expected size upon 2% agarose gel electrophoresis stained, with GelRed nucleic acid staining (Biotium, USA).

5.3.6 Quantitative real-time PCR

qPCR experiments were performed in an iQ5 real-time thermal cycler (Bio-Rad), using EvaGreen® Supermix (Bio-Rad). Each 15 µl reaction comprised 5 µl of the diluted cDNA, 7.5 µl EvaGreen Supermix, 200 nM of each primer and 1.7 µl of sterile distilled water. Thermal cycling for all genes was carried out under the following conditions: 3 min. of polymerase activation at 95°C; followed by 45 cycles of denaturation at 95°C for 10 s, and 30 s annealing at the annealing temperature for each gene (**Table 5.2**). A melting curve analysis was performed at the end of the PCR run over the range 55-95°C, increasing the temperature in a stepwise fashion by 0.5°C every 10s. Each set of reactions included a negative control with no template. Dissociation curves and agarose gel electrophoresis were used to analyze non-specific PCR products. The efficiency of each primer pair (amplification efficiency, E) was experimentally tested with the LinRegPCR program (Ramakers *et al.*, 2003), which uses a linear regression analysis of fluorescence data from the exponential phase of PCR amplification to determine amplification efficiency (E). Two independent experiments comprising three biological replicates and two technical replicates were used for each sample.

5.3.7 Assessment of gene expression stability

The expression stability of each candidate RG and the best combination of RGs were obtained using a pairwise method by geNorm (Vandesompele *et al.*, 2002) and a model-based method by NormFinder (Andersen *et al.*, 2004) software. The geNorm algorithm calculates the expression stability (M) based on the average pairwise variation between all RGs tested. The gene with the lowest M value is considered to have the most stable expression, while that with the highest M value presents a high variance in its

expression (Vandesompele *et al.*, 2002). Moreover, geNorm estimates the normalization factor (NF) using the geometric mean of expression levels of n best RGs, using a pairwise variation (V) with a cut-off value of 0.15. By contrast, NormFinder uses a model-based algorithm that takes into account the overall stability, as well as the stability of any groups that may be present in the sample set. This software ranks the stability values of candidate RGs, being the lowest stability value to the most stable expressed gene (Andersen *et al.*, 2004).

The analyses were performed considering two different datasets: i) all *C. arabica* - *C. kahawae* samples; and ii) all *C. kahawae* samples. Moreover, in order to test for differences between *C. kahawae* aggressiveness patterns, the *C. arabica* - *C. kahawae* samples were also analyzed separately according to the different isolates under study (Ang29, Ang67, Zim12 and Que2).

For *C. arabica* - *C. kahawae* samples a biomass correction step was required. This method, previously described for *Hemileia vastatrix* (Vieira *et al.*, 2011), takes into account the variation of fungal biomass within the samples across the infection process. Therefore, DNA was used to estimate the fungal biomass across key time points of pathogenesis. The DNA was extracted, from the same hypocotyls used in RNA assays, using the DNeasy Plant mini kit (Qiagen) as recommended by the manufacturer. Genomic DNA purity, concentration and integrity were determined as described above for RNA samples, and samples were diluted to 10ng.μl^{-1}.

qPCR amplifications were performed using the same conditions previously described in section "*Quantitative real-time PCR*", with the primer set *ck39066* (homologue to hypothetical protein CGGC5_4189) (**Table 5.2)**. The Cq results were then used to normalize the Cq of RGs using the following formula:

$$Cq\ corrected = Cq\ cDNA\ reference\ genes - CqDNAck39066$$

It should be noted that this additional correction step is only required for the selection of stable RGs, not in the subsequent gene expression analysis. The Cq value of each RG obtained for all samples was transformed into relative quantity (Q) as compared to the Cq value of ungerminated conidia samples, using the formula previously described by (Pfaffl, 2001). The definition of the optimal number of genes required for normalization

was conducted by geNorm pairwise variation analysis (Vandesompele *et al.*, 2002). A comprehensive ranking was established by calculating the arithmetic mean ranking value of each gene using the two applets, and each gene was ranked from 1 (most stable) to 8 (least stable) (Wang *et al.*, 2012).

5.3.8 Expression profiles of two genes of interest

The expression profile of two pathogenesis-related genes, previously described as being involved in host penetration of several fungi, was analyzed in this work to validate the RGs selected for each dataset. The target genes *cat2* and *thr1* encode a bifunctional catalase-peroxidase, involved in ROS detoxification (Bourdais *et al.*, 2012; Brown *et al.,* 2008; Tanabe *et al.,* 2011; Yarden *et al.*, 2014), and a trihydroxynaphthalene reductase, involved in the melanin biosynthesis pathway (Bourdais *et al*., 2012; Perpetua et al., 1996; Thompson *et al.,* 2000; Tsuji *et al.*, 2003), respectively. *C. kahawae* genes homologous to *cat2* and *thr1* were retrieved from the *C. kahawae* RNA-seq database, showing differential expression in the early phases of the infection process (unpublished data). Specific primers (**Table 5.2**) were designed with PerlPrimer v1.1.17 (Marshall, 2004), as described above.

To assess gene expression, relative quantities (RQ) were calculated for both (RGs and genes of interest) using the formula $RQ = E^{\Delta Cq}$ where E represents the amplification efficiency (E) for each gene and ΔCq the difference between the Cq from each target sample and the conidia sample ($\Delta Cq = Cq\,conidia - Cq\,target$) (Pfaffl, 2001). A normalization factor (NF), calculated as the geometric mean of the relative quantities of the RGs selected for each normalization approach was used to obtain the normalized relative quantities (Pfaffl, 2001). In this work, six different normalization factors were tested: **i**) best three ranking genes selected for *C. arabica* - *C. kahawae* samples (**NF Global** *C. arabica* - *C. kahawae* interaction); **ii**) best normalization factor identified by geNorm for both datasets (**NF Best geNorm**); **iii**) best three ranking genes selected for *C. arabica* - *C. kahawae* samples according to the different *C. kahawae* isolates under study (**NF Best aggressiveness pattern**); **iv**) best two ranking genes common to both datasets (**NF Best two ranking genes**); **v)** best three ranking genes selected for *C. kahawae* samples (**NF Global *C. kahawae***); **vi**) worst ranking genes selected for both datasets (**NF Worst**). Statistically significant differences between the six normalization

factors used were determined by the Kruskall-Wallis test. Potential differences between the time points for *C.arabica* - *C.kahawae* and between lifecycle stages for *C. kahawae* were tested using the Mann-Whitney test. Both tests were computed in IBM® SPSS® Statistics version 20.0.0 (SPSS Inc., USA) software and they were considered significant when $p < 0.05$.

5.4 Results and discussion

Reference gene validation for qPCR expression studies has become a fundamental requisite for reliable quantification results. Here, we describe an assessment of eight candidate RGs for their use as internal controls in gene expression studies of *C. kahawae* and *C. arabica* - *C. kahawae* interaction samples. In this work, two different approaches were applied to identify suitable candidate RGs: analysis of traditional RGs previously validated for other fungi (literature); and of new candidate RGs from Illumina RNA-seq datasets. The genes selected from the literature included *Act*, *Cyp*, and PP1, while the genes selected from Illumina RNA-seq datasets were *ck20430*, *ck28444*, *ck36020*, *ck48742* and *ck34620* (**Table 5.2**).

5.4.1 Amplification specificity and efficiency

To investigate the expression stability of the candidate RGs selected, transcript levels of the eight candidates were measured by qPCR using gene-specific primer pairs (**Table 5.2**). Specificity of real-time PCR products was confirmed by the presence of a single peak in the melting curve and the presence of a single band with the expected size after agarose gel electrophoresis (**Figure A4.1**). No amplification was obtained for the negative control. The average PCR efficiency of primers ranged from 1.902 to 1.950 (**Table 5.2**) as calculated by LinRegPCR (Ramakers *et al.*, 2003). Considering that the efficiency value for one primer pair could be different among the type of samples (*C. kahawae* samples or *C. arabica* - *C. kahawae* samples), a separate analysis was performed, but no significant differences were observed (**Table A4.1**).

5.4.2 Determination of Cq values and variation on candidate Rgs

Following PCR amplification, a general overview of the expression profile and relative abundance of each candidate gene was obtained by plotting the Cq values obtained for *C. arabica* - *C. kahawae* samples as well as for *C. kahawae* samples (**Figure 5.1**).

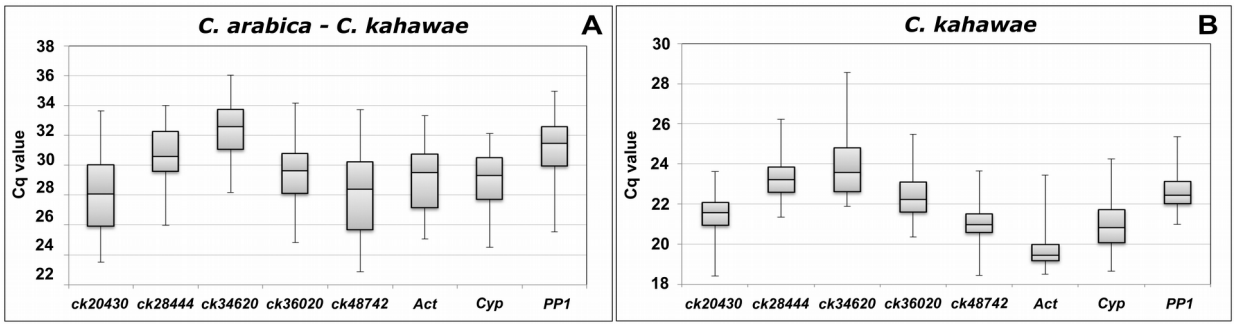


Figure 5.1 - Box and whisker plots of Cq values for each reference gene across the experimental samples. **A)** *C. arabica-C. kahawae* samples; **B**) *C. kahawae* samples. The boxes indicate the 25th and 75th percentiles. Lines within the boxes represent the median Cq values; the whiskers mark minimum and maximum values in each data set.

Most candidate RGs displayed median Cq values for fungus samples ranging from 19 to 23, indicating a moderately high level of expression. However, in *C. arabica* - *C. kahawae* samples, the median Cq for the same candidate RGs ranged from 28 to 32. The minimum Cq values, meaning higher abundance, were observed for *Act* (18.5) in mature appressoria formed after 18-22 h on polystyrene, while *ck34620* had the lowest expression value (35.5) on the early stages of the infection process. Overall, gene expression variation across samples taken from different stages of the infection process ranged from 7.6 to 10.9 Cq, while for *C. kahawae* samples ranged only between 4.4 and 6.7 Cqs (**Figure 5.1**). These changes in Cq values may reflect the variations in the amount of fungal RNA present in *C. arabica* - *C. kahawae* interaction samples at each time point. According to previous studies (Vieira *et al.*, 2011), the effect of fungal biomass variation in *C. arabica* - *C. kahawae* samples can be corrected using the Cq value of the amplification of a gene from DNA extracted from the same samples used for RNA quantification. In spite of this, occasionally, the Cq value for each candidate RG did not followed a strictly parallel line to that of genomic DNA, suggesting variations in the expression levels of candidate RGs (**Figure A4.2**). Though preliminary information can be obtained through absolute Cq analysis (**Figure 5.1**), to correctly assess the

expression stability of candidate genes, it is crucial to use statistical tools like geNorm and NormFinder to determine the best set of RGs.

However, since these statistical tools were developed to assess the stability of RGs when the amount of RNA is similar under studied conditions (Andersen *et al.*, 2004; Vandesompele *et al.*, 2002), a correction step was applied to allow an accurate analysis of candidate RGs stability in *C. arabica - C. kahawae* samples.

5.4.3 Analysis of gene expression stability data

The expression stability of the candidate RGs, as well as the number of RGs necessary for accurate gene-expression profiling, was performed using two different statistical applets, geNorm (Vandesompele *et al.*, 2002) and NormFinder. Though both aim to determine which candidate RGs are the most stable under certain conditions, they run under different algorithms and mathematical models. Therefore, the stability ranking of the putative RGs might differ depending on the software used as previously described (Borges *et al.*, 2014; Figueiredo *et al.*, 2013). The results were analyzed dividing the data into two different datasets: i) all *C. arabica - C. kahawae* samples; ii) all *C. kahawae* samples. Moreover, the datasets of *C. arabica - C. kahawae* samples were analyzed separately considering the different fungal isolates under study (Ang29, Ang67, Zim12 and Que2) to check if the set of candidate RGs changes according to the interaction established.

As shown in **Table 5.3** slight differences were found between the stability rankings of RGs provided by the two computational programs for the *C. arabica - C. kahawae* samples, being the *PP1*, *ck34620* and *Act* selected as the most stable genes. By contrast, in the analysis of *C. kahawae* samples, the best three selected RGs changed according to the software used. In the geNorm analysis, the best set of candidate RGs were *ck20430*, *ck48742* and *PP1*, while in the NormFinder analysis the selected set of RGs were *PP1*, *Act* and *ck20430*, followed by *ck48742*. This incongruence was previously described in other RG validation studies (Borges *et al.*, 2014; Figueiredo *et al.*, 2013; Huis *et al.*, 2010) and reveals the importance of using a comprehensive ranking analysis. Currently several strategies exist to create a comprehensive stability ranking which integrate the results of the two software. Here, a certain weight was

assigned to each gene corresponding to the rank obtained from each program (e.g. 1-most stable to 8-least stable). Subsequently, the rank aggregation relied on straightforward arithmetic and geometric means of the ranks (**Table 5.3**) (Wang *et al.*, 2012).

Table 5.3 - Comprehensive ranking of candidate reference genes for each of the datasets used: **i**) all *C. arabica-C. kahawae* interaction samples**; ii)** all *C. kahawae* samples

Gene Name	Global *C. arabica-C. kahawae*					Global *C. kahawae*				
	NormFinder		geNorm		Overall ranking	NormFinder		geNorm		Overall ranking
	Stability value	Rank	M value	Rank		Stability value	Rank	M value	Rank	
ck48742	0.66	8	1.68	7	8	0.37	4	0.51	1	3
ck20430	0.62	7	1.61	6	7	0.34	3	0.51	1	2
ck36020	0.57	6	1.48	5	6	0.60	8	1.36	7	8
ck28444	0.55	5	1.37	3	4	0.38	5	0.99	4	5
Cyp	0.53	4	1.42	4	4	0.39	6	1.05	5	6
Act	0.51	3	1.15	1	3	0.26	2	0.89	3	3
ck34620	0.43	2	1.15	1	1	0.47	7	1.17	6	7
PP1	0.27	1	1.28	2	1	0.24	1	0.78	2	1

Taking into account the comprehensive ranking results, the top three most stable genes for *C. arabica - C. kahawae* samples were *PP1*, *ck34620* and *Act*, while for *C. kahawae* samples were *PP1*, *Act* and *ck20430* (**Table 5.4**). Although the best set of RGs changed according to the type of samples, *PP1* and *Act* seemed to be stable under all tested conditions. The higher stability of *Act* and *PP1* was previously described for *Beauveria bassiana* under all analyzed conditions (Zhou *et al.*, 2012). However, for other fungi such as *Metarhizium anisopliae*, *Act* was not among the most stable RGs (Fang and Bidochka, 2006) which reinforce the importance of developing these studies.

The analysis carried out by geNorm enabled the determination of the optimal number of RGs through the calculation of pairwise variation (Vn/Vn+1) between two sequential candidate RGs. High values indicate the need for the inclusion of another gene to obtain a reliable normalization factor, which should contain at least two RGs. Thus, extra RGs can be included until the Vn/Vn+1 is smaller than a threshold of 0.15, as recommended (Vandesompele *et al.*, 2002). However, this is not an absolute value and it can change according to the data (Vandesompele *et al.*, 2002). As shown in **Figure 5.2**, the pairwise variation values, for both datasets, are higher than the recommended cut-off value.

Table 5.4 - Normalization factors tested for gene expression analysis referring to the candidate reference genes included for each sample type

Normalization factors	*C. arabica-C. kahawae*				*C. kahawae*
[1]**NF Global** *C. arabica-C. kahawae*	PP1; Act; ck34620				PP1; Act; ck34620
[2]**NF Best geNorm**	PP1; Act; ck34620; ck28444; Cyp				ck48742; Act; PP1; ck20430
[3]**NF Two Best ranking genes**	PP1; Act				PP1; Act
[4]**NF Global** *C. kahawae*	PP1; Act; ck20430				PP1; Act; ck20430
[5]**NF Worst**	ck20430; ck48742; ck36020				ck34620; ck36020
[6]**NF Best aggressiveness pattern**	**Ang29**	**Ang67**	**Zim12**	**Que2**	
	Cyp; Act; ck28444	Cyp; Act; ck34620	Cyp; Act; PP1	Cyp; Act; PP1	

[1] The best three ranking genes selected for *C. arabica - C. kahawae* samples;
[2] The best normalization factor identified by geNorm for both datasets;
[3] The best two ranking genes common to both datasets
[4] The best three ranking genes selected for C. *kahawae* samples
[5] The worst ranking genes selected for both datasets
[6]The best three ranking genes selected for *C. arabica - C. kahawae* samples according to the different isolates under study

Based on this parameter, the use of five reference genes (V5/6= 0.224248) for the *C. arabica - C. kahawae* sample dataset and four (V4/5 = 0.207316) for the fungal dataset seems to be the best approach (lowest value). Therefore, the **NF Best geNorm** for *C. arabica - C. kahawae* samples comprises the use of *PP1*; *Act*; *ck34620*; *ck28444*; *Cyp* while the **NF Best geNorm** for *C. kahawae* samples includes *ck48742*; *Act*; *PP1*; *ck20430* as best reference genes. However it is crucial to validate if the use of five/four genes is indeed necessary to obtain a proper data normalization, or if this purpose can be reached with only two or three RGs, since a trade-off between the practical considerations, such as time/cost, and accuracy, should be considered, especially if no significant differences are observed. A separate analysis of the *C. arabica - C. kahawae* samples according to the different isolates under study was further performed, in order to test for significant differences related with aggressiveness patterns. Despite the slight differences between the results provided by geNorm and NormFinder, the same best three RGs (*Act*, *Cyp* and *PP1*) were identified for Zim12 and Que2, while for Ang29 the best set was *Act*, *Cyp* and *ck28444*, and for Ang67 *Act*, *Cyp* and *ck34620*, being *PP1* the fourth most stable RG for the latter isolate (**Table A4.2**). Although these results point to a different conclusion from the global approach, almost all genes showed a stability

value lower than the recommended cut-off value of 1.5 (Vandesompele *et al.*, 2002), noting their high stability. However, a proper validation of the selected candidate RGs will be crucial to prove if these differences on the Normalization factor used can significantly change the expression of target genes.

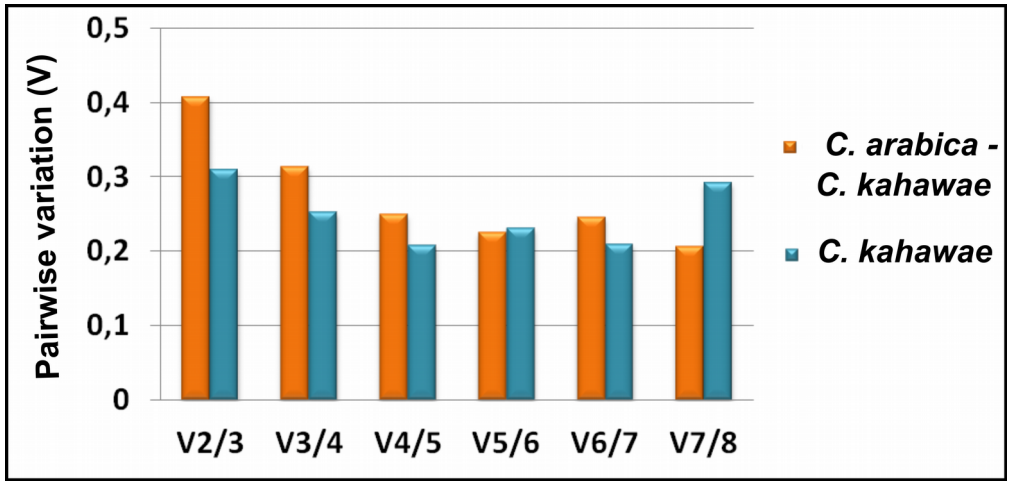


Figure 5.2 - Prediction of the optimal number of reference genes required for effective normalization. Pairwise variation (V) of the candidate reference genes calculated by geNorm using the two different datasets studied: **i)** all *C. arabica-C. kahawae* samples; **ii)** all *C. kahawae* samples.

The worst candidate RGs for *C. kahawae* samples were *ck34620* and *ck36020*, while for *C. arabica - C. kahawae* samples were *ck36020*, *ck48742* and *ck20430*. Although all least stable genes were candidates retrieved from the RNA-seq datasets, these are nevertheless a promising source of additional (and potentially independent) qPCR reference genes, since, for in both types of samples (*C. kahawae* and *C. arabica - C. kahawae* interactions), the best set of reference genes (**Table 5.3 and Table 5.4**) included genes from the literature and genes from RNA-seq data. Similar results were observed in a previous study involving *Fusarium graminearum* (Kim and Yun, 2011). Moreover, the current RNA-seq dataset is derived from *C. kahawae* infected coffee leaves, therefore comprising a mixture of plant and fungal RNA which restricts the number of pathogen genes retrieved. Still, this dataset was sufficient to generate five candidate RGs, suggesting that larger datasets might provide higher numbers of candidate RGs.

5.4.4 Expression analysis of pathogenesis-related genes

Two different genes of interest (*cat2* and *thr1*), related with the early stages of the infection process, were used to validate the best normalization factor for qPCR data analysis of *C. kahawae*, regarding its aggressiveness pattern or developmental stages, using the conidia as control.

Six different normalization factors were tested, according to the results obtained by geNorm and NormFinder, in order to correctly choose the best set of reference genes as shown in **Table 5.4: i) NF Global** *C. arabica - C. kahawae* interaction; **ii) NF Best geNorm**; **iii) NF Best aggressiveness pattern**; **iv) NF Best two ranking genes**; v) **NF Global** *C. kahawae*; **vi) NF Worst**. Considering the normalization factors tested, the expression profile of *thr1* showed only slight differences between most of them (**Figures A4.3 and A4.3**), being the worst normalization factor the exception. For the *C. arabica - C. kahawae* dataset, no significant differences were observed between the first five NFs, when the Kruskal-Wallis test was applied (**Table A4.3**). However, significant differences were observed between these and the worst normalization factor (**Table A4.3**). For the fungal datasets a similar result were observed (**Table A4.4**). The expression profile of *cat2* only changed drastically when the worst normalization factor was applied (**Figures A4.3 and A4.4**), but significant statistical differences were observed between the different normalization factors tested (**Table A4.5**).

When applying **NF Best geNorm**, which includes the five best genes selected for the *C. arabica - C. kahawae* dataset, as well as the four best genes selected for the *C. kahawae* dataset, no significant differences in expression levels were detected in comparison with the use of **NF Global** *C. arabica - C. kahawae* or **NF Globa**l *C. kahawae* (**Tables A4.5 and A4.6**), with only 3 reference genes. Thus, adding two or one more reference genes did not increased the accuracy of the results. Moreover, no significant differences were observed when comparing **NF Best aggressiveness pattern** with **NF Global** *C. arabica - C. kahawae* or **NF Best geNorm** (**Table A4.5**). These results show that the same normalization factor could be used to normalize all the isolates regardless of its aggressiveness pattern. In contrast, the best two ranking genes seem to have significant differences, when compared to **NF Best geNorm**, **NF Best aggressiveness pattern** or **NF Global** *C. arabica - C. kahawae* (**Table A4.5**).

These results shows that, although *PP1* and *Act* genes appear to be stable under all conditions studied, the use of only these two RGs is not the best approach for an accurate normalization approach. Finally, the worst normalization factor has significant differences when compared with the five normalization factors previously described for both datasets (**Tables A4.5 and A4.6**).

In summary, under the present experimental system, the best normalization factor for *C. arabica - C. kahawae* interaction samples comprises the use of *PP1*, *Act* and *ck34620*, while for *C. kahawae* samples the best normalization factor is provided by the joint use of *PP1*, *Act* and *ck20430*. Moreover, the significant differences observed on the expression profile of the genes of interest when normalized with the Best or the Worst NF shows the need to validate the best set of reference genes for each specific experimental condition (**Figure 5.3** and **Figure 5.4**). Such, sample-specific normalization requisites were previously described for many other organisms, namely *Hemileia vastatrix, Fusarium graminearum, Vitis vinifera and Caragana intermedia* (Borges *et al.*, 2014; Kim and Yun, 2011; Vieira *et al.*, 2011; Zhu *et al.*, 2008).

The expression profiles of *cat2* and *thr1* during the infection time-course (*C. arabica - C. kahawae* samples) were similar, showing a highest expression peak in the early stages of the infection process (24/48h) followed by a slight decrease over time (**Figure 5.3** and **Figure 5.4**). However, significant differences on the expression between the time points were only statistical significant for the Ang 67 isolate (**Tables A4.7 and A4.8**). In contrast, the expression profile of *Cat2* for Zim 12 isolate had a significant slight increase on the expression over the time (**Figure 5.4, Tables A4.7 and A4.8**). Previous studies on *Colletotrichum acutatum* showed a high level of expression for *cat2* during appressoria formation when compared with mycelia grown under nitrogen limitation or a complete nutrient supply (Brown *et al.*, 2008).

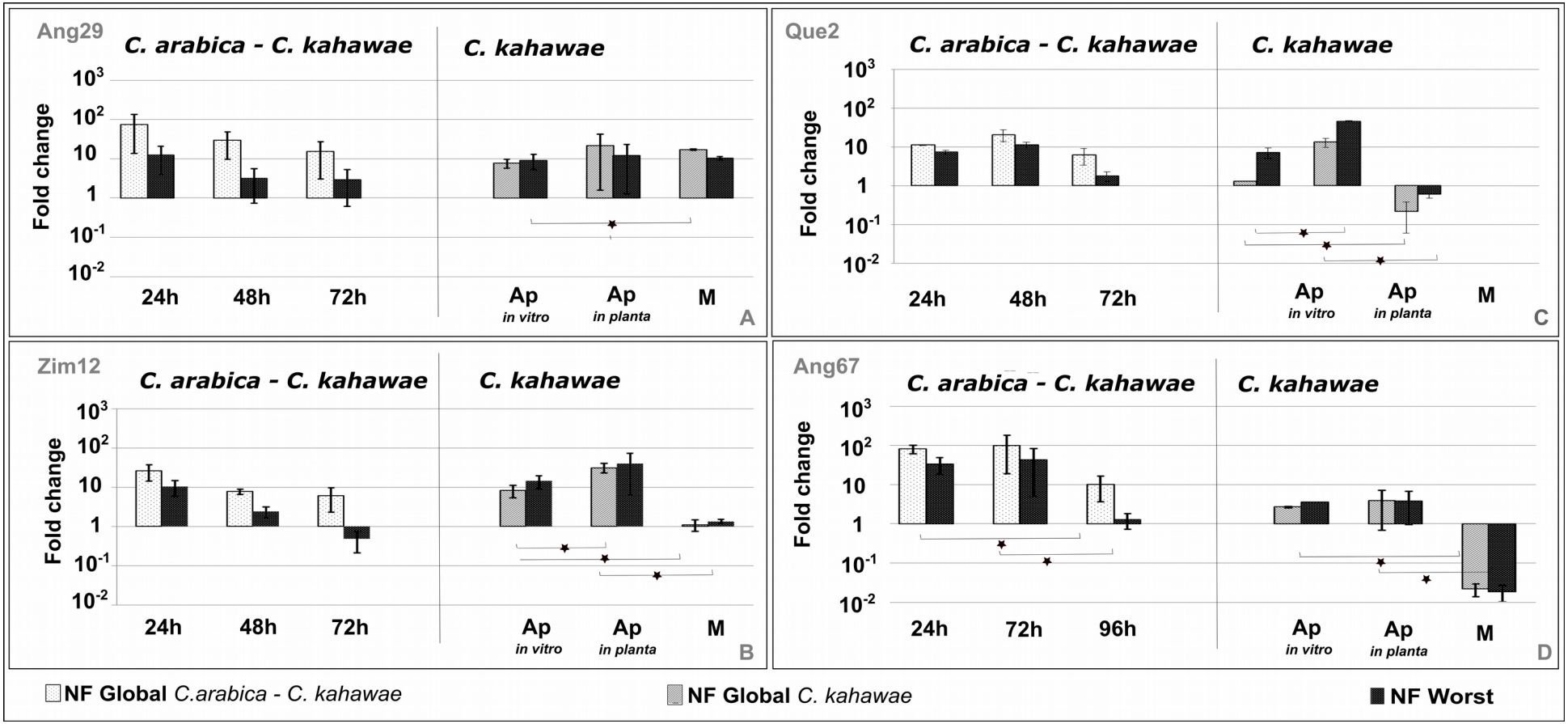


Figure 5.3 - Relative quantification of *thr1* expression using the Best and the Worst normalization factors (NF). Expression profiles are presented per isolate (Ang29 (A), Zim 12 (B), Que2 (C) and Ang67 (D)), during the early stages of infection process and growth (Ap: Appressoria; M: Mycelium). The *C. arabica – C. kahawae* samples were normalized with **NF Global** *C. arabica - C. kahawae* interaction (*PP1; Act; ck34620*) and **NF Worst** (*ck20430; ck48742; ck36020*), while the *C. kahawae* samples were normalized with **NF Global** *C. Kahawae* (*PP1; Act; ck20430*) and **NF Worst** (*ck34620; ck36020*).

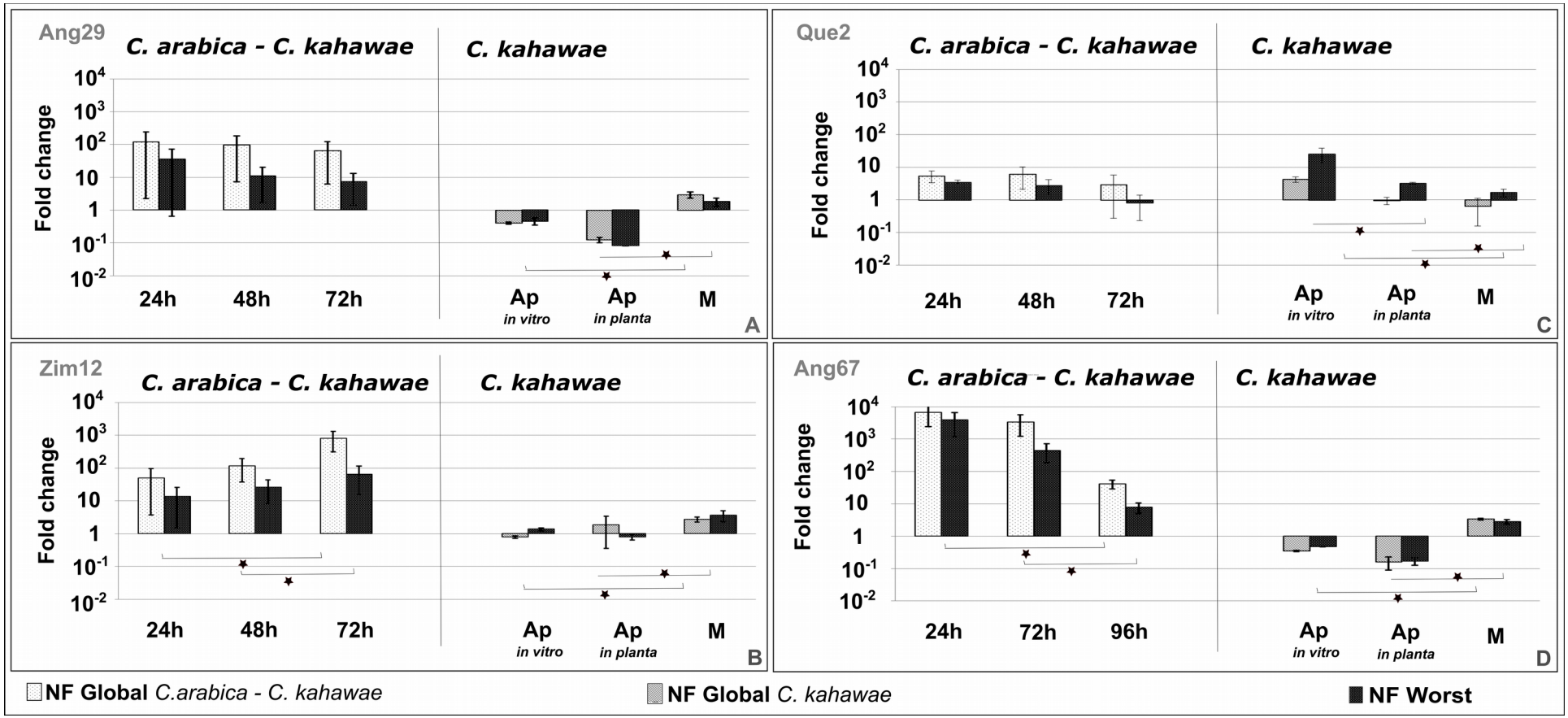


Figure 5.4 - Relative quantification of *cat2* expression using the best and the worst normalization factors (NF). Expression profiles are presented per isolate (Ang29 (A), Zim 12 (B), Que2 (C) and Ang67 (D)), during the early stages of infection process and growth (Ap: Appressoria; M: Mycelium). The *C. arabica – C. kahawae* samples were normalized with **NF Global** *C. arabica - C. kahawae* interaction (*PP1; Act; ck34620*) and **NF Worst** (*ck20430; ck48742; ck36020*), while the *C. kahawae* samples were normalized with **NF Global** *C. Kahawae* (*PP1; Act; ck20430*) and **NF Worst** (*ck34620; ck36020*).

On the other hand, for the *C. kahawae* samples, different expression profiles were observed between the two genes. *cat2* seemed to be expressed only in the mycelium and repressed in appressoria, both in planta and in vitro, with significant differences (**Figure 5.4**), while *thr1* expression was higher in planta and in vitro appressoria and was repressed on mycellium for almost all isolates, with significant differences (**Figure 5.3**).

For Ang 29 the thr1 gene seems to be highly expressed in all *C. kahawae* samples with only significant difference between Ap in vitro and mycellium (**Tables A4.7 and A4.8**). Previous studies on *Colletotrichum lagenarium* showed an increase on the expression of *thr1* during spore germination (Perpetua *et al.*, 1996). Despite the interesting expression profile of these genes, subsequent targeted expression studies will be required to associate different expression profiles with the aggressiveness patterns of the isolates.

5.5 Conclusions

In the present study, we evaluated the expression stability of eight candidate reference genes across several *C. kahawae* samples representing different growth and infection stages, with the aim of identifying the best set for data normalization of gene expression studies. New candidate reference genes were selected based on a genome wide approach (RNA-seq datasets) and among them most had expression stability similar to the typical reference genes selected in the literature. This highlights RNA-seq data as alternative sources of reference genes, in addition to classical (i.e., literature-based) reference genes. Two main normalization factors were selected according to the type of samples under study, applying a combination of reference genes: *PP1*, *Act* and *ck34620* were the best set of reference genes when *C. arabica - C. kahawae* samples were used, while *PP1*, *Act* and *ck20430* were the best set when only *C. kahawae* samples were used. Unfortunately, although *PP1* and *Act* were the two most stably expressed RGs, the NF common to both type of samples, relying only on these RGs, does not seem to constitute a good normalization factor for all tested samples. The expression profiles of *cat2* and *thr1* during the infection time-course on *C. arabica - C. kahawae* samples were similar, with the highest expression peak in the early stages of

the infection process 24/48 hpi and a slight decrease over time. For *C. kahawae* samples, different expression profiles were observed between the two genes, being *cat2* more expressed in the mycelium and repressed in appressoria, while *thr1* expression was higher *in planta* and *in vitro* appressoria. This work provides the first reference genes specifically established for *C. kahawae* samples, and this information will greatly facilitate future studies of gene expression in *C. kahawae*.

5.6 References

Andersen, C.L., Jensen, J.L. and Ørntoft, T.F. (2004) Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. *Cancer Res* **64**, 5245-5250.

Borges, A.F., Fonseca, C., Ferreira, R.B., Lourenço, A.M. and Monteiro, S. (2014) Reference gene validation for quantitative RT-PCR during biotic and abiotic stresses in *Vitis vinifera*. *PLoS One* **9**, e111399.

Bourdais, A., Bidard, F., Zickler, D., Berteaux-Lecellier, V., Silar, P. and Espagne, E. (2012) Wood utilization is dependent on catalase activities in the filamentous fungus *podospora anserina*. *PLoS One* **7**.

Brown, S.H., Yarden, O., Gollop, N., Chen, S., Zveibil, A., Belausov, E. and Freeman, S. (2008) Differential protein expression in *Colletotrichum acutatum*: Changes associated with reactive oxygen species and nitrogen starvation implicated in pathogenicity on strawberry. *Mol. Plant Pathol.* **9**, 171–190.

Bustin, S. a, Beaulieu, J.-F., Huggett, J., Jaggi, R., Kibenge, F.S.B., Olsvik, P. a, Penning, L.C. and Toegel, S. (2010) MIQE précis: Practical implementation of minimum standard guidelines for fluorescence-based quantitative real-time PCR experiments. *BMC Mol. Biol.* **11**, 74.

Coker, J. and Davies, E. (2003) Selection of candidate housekeeping controls in tomato plants using EST data. *BioTechniquesniques* **35**, 740–748.

Fang, W. and Bidochka, M.J. (2006) Expression of genes involved in germination, conidiogenesis and pathogenesis in *Metarhizium anisopliae* using quantitative real-time RT-PCR. *Mycol. Res.* **110**, 1165–71.

Figueiredo, A., Loureiro, A., Batista, D., Monteiro, F., Várzea, V., Pais, M.S., Gichuru, E.K. and Silva, M.C. (2013) Validation of reference genes for normalization of qPCR gene expression data from *Coffea* spp. hypocotyls inoculated with *Colletotrichum kahawae*. *BMC Res. Notes* **6**, 388.

Fino, J., Figueiredo, A., Loureiro, A., Gichuru, E.K. and Várzea, V. (2014) Transcriptional profiling of compatible and incompatible *Coffee* - *Colletotrichum kahawae* interactions through RNA-Seq analysis. *25th Int. Conf. Coffee Sci.*, 42–46.

Gichuru, E.K., Agwanda, C.O., Combes, M.C., Mutitu, E.W., Ngugi, E.C.K., Bertrand, B. and Lashermes, P. (2008) Identification of molecular markers linked to a gene conferring resistance to coffee berry disease (*Colletotrichum kahawae*) in *Coffea arabica*. *Plant Pathol.* **57**, 1117–1124.

Hindorf, H. and Omondi, C.O. (2011) A review of three major fungal diseases of *Coffea arabica* L. in the rainforests of Ethiopia and progress in breeding for resistance in Kenya. *J. Adv. Res.* **2**, 109–120.

Hruz, T., Wyss, M., Docquier, M., *et al*. (2011) RefGenes: identification of reliable and condition specific reference genes for RT-qPCR data normalization. *BMC Genomics* **12**, 156.

Huggett, J., Dheda, K., Bustin, S. and Zumla, A. (2005) Real-time RT-PCR normalisation; strategies and considerations. *Genes Immun* **6**, 279–284.

Huis, R., Hawkins, S. and Neutelings, G. (2010) Selection of reference genes for quantitative gene expression normalization in flax (*Linum usitatissimum* L.). *BMC Plant Biol.* **10**, 71.

Kim, H.K. and Yun, S.H. (2011) Evaluation of potential reference genes for quantitative RT-PCR analysis in *Fusarium graminearum* under different culture conditions. *Plant Pathol. J.* **27**, 301–309.

Kleemann, J., Takahara, H., Stueber, K. and O'Connell, R. (2008) Identification of soluble secreted proteins from appressoria of *Colletotrichum higginsianum* by analysis of expressed sequence tags. *Microbiology* **154**, 1204– 1217.

Lin, F., Jiang, L., Liu, Y., Lv, Y., Dai, H. and Zhao, H. (2014) Genome-wide identification of housekeeping genes in maize. *Plant Mol. Biol.* **86**, 543–554.

Llanos, A., François, J.M. and Parrou, J. (2015) Tracking the best reference genes for RT-qPCR data normalization in filamentous fungi. *BMC Genomics* **16**:71

Loureiro, A., Azinheira, H.G., Silva, M. do C. and Talhinhas, P. (2015) A method for obtaining RNA from *Hemileia vastatrix* appressoria produced *in planta*, suitable for transcriptomic analyses. *Fungal Biol.* **119**, 1093–1099.

Loureiro, A., Nicole, M.R., Várzea, V., Moncada, P., Bertrand, B. and Silva, M.C. (2012) Coffee resistance to *Colletotrichum kahawae* is associated with lignification,

accumulation of phenols and cell death at infection sites. *Physiol. Mol. Plant Pathol.* **77**, 23–32.

Marshall, O. (2004) PerlPrimer: cross-platform, graphical primer design for standard, bisulphite and real-time PCR. *Bioinformatics* **20**, 2471–2472.

Montecinos, A.E., Guillemin, M.L., Couceiro, L., Peters, A.F., Stoeckel, S. and Valero, M. (2017) Hybridization between two cryptic filamentous brown seaweeds along the shore: analysing pre- and postzygotic barriers in populations of individuals with varying ploidy levels. *Mol. Ecol.*, 1–16.

Perpetua, N.S., Kubo, Y., Yasuda, N., Takano, Y. and Furusawa, I. (1996) Cloning and characterization of a melanin biosynthetic *THR1* reductase gene essential for appressorial penetration of *Colletotrichum lagenarium*. *Mol. Plant. Microbe. Interact.* **9**, 323–329.

Pfaffl, M. (2001) A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res.* **29**, e45.

Ramakers, C., Ruijter, J.M., Lekanne Deprez, R.H. and Moorman, A.F.M. (2003) Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. *Neurosci. Lett.* **339**, 62–66.

Silva, C., Várzea, V., Guerra-guimarães, L., Azinheira, H.G., Fernandez, D., Petitot, A., Bertrand, B., Lashermes, P. and Nicole, M. (2006) Coffee resistance to the main diseases: leaf rust and coffee berry disease. *Braz. J. Plant Physiol.* **18**, 119–147.

Silva, M., Nicole, M., Rijo, L., Geiger, J. and Rodrigues Jr., C. (1999) Cytochemical aspects of the plant–rust fungus interface during the compatible interaction *Coffea arabica* (cv. Caturra) – *Hemileia vastatrix* (race III). *Int. J. Plant Sci.* **160**, 79–91.

Tanabe, S., Ishii-Minami, N., Saitoh, K.I., Otake, Y., Kaku, H., Shibuya, N., Nishizawa, Y. and Minami, E. (2011) The role of catalase-peroxidase secreted by *Magnaporthe oryzae* during early infection of rice cells. *Mol. Plant. Microbe. Interact.* **24**, 163–171.

Thompson, J.E., Fahnestock, S., Farrall, L., Liao, D.I., Valent, B. and Jordan, D.B. (2000) The second naphthol reductase of fungal melanin biosynthesis in *Magnaporthe grisea*. *J. Biol. Chem.* **275**, 34867–34872.

Tsuji, G., Sugahara, T., Fujii, I., Mori, Y., Ebizuka, Y., Shiraishi, T. and Kubo, Y. (2003) Evidence for involvement of two naphthol reductases in the first reduction step of melanin biosynthesis pathway of *Colletotrichum lagenarium*. *Mycol. Res.* **107**, 854–860.

Vandesompele, J., Preter, K. De, Pattyn, F., Poppe, B., Roy, N. Van, Paepe, A. De and Speleman, F. (2002) Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.* **3**, RESEARCH0034.

Velculescu, V., Madden, S., Zhang, L., Lash, A., Yu, J., Rago, C. and Al., E. (1999) Analysis of human transcriptomes. *Nat. Genet.* **23**, 387–388.

Vieira, A., Talhinhas, P., Loureiro, A., Duplessis, S., Fernandez, D., Silva, M.D.C., Paulo, O.S. and Azinheira, H.G. (2011) Validation of RT-qPCR reference genes for *in planta* expression studies in *Hemileia vastatrix*, the causal agent of coffee leaf rust. *Fungal Biol.* **115**, 891–901.

Vossen, H.A.M. van der, Cook, R.T.A. and Murakaru, G.N.W. (1976) Breeding for resistance to coffee berry disease caused by *Colletrotrichum caffeanum* Noack (sensu hindorf) in *Coffea arabica* L. I. Methods of preselection for resistance. *Euphytica* **25**, 733–745.

Wang, Q., Ishikawa, T., Michiue, T., Zhu, B.L., Guan, D.W. and Maeda, H. (2012) Stability of endogenous reference genes in postmortem human brains for normalization of quantitative real-time PCR data: Comprehensive evaluation using geNorm, NormFinder, and BestKeeper. *Int. J. Legal Med.* **126**, 943–952.

Wong, M. and Medrano, J. (2005) Real-time PCR for mRNA quantitation. *Biotechniques* **39**, 75–85.

Yarden, O., Veluchamy, S., Dickman, M.B. and Kabbage, M. (2014) Sclerotinia sclerotiorum catalase *SCAT1* affects oxidative stress tolerance, regulates ergosterol levels and controls pathogenic development. *Physiol. Mol. Plant Pathol.* **85**, 34–41.

Zhou, Y.-H., Zhang, Y.-J., Luo, Z.-B., Fan, Y.-H., Tang, G.-R., Liu, L.-J. and Pei, Y. (2012) Selection of optimal reference genes for expression analysis in the entomopathogenic fungus *Beauveria bassiana* during development, under changing nutrient conditions, and after exposure to abiotic stresses. *Appl. Microbiol. Biotechnol.* **93**, 679–85.

Zhu, J., He, F., Song, S., Wang, J. and Yu, J. (2008) How many human genes can be defined as housekeeping with current expression data? *BMC Genomics* **9**, 172.

Final remarks

6

6.1 General overview

Plant pathogens are still emerging and will continue to do so in agro-ecosytems throughout the world (McDonald and Stukenbrock, 2016). Fungi (*sensu lato*) are responsible for ~30% of the emerging diseases in plants and they can impact negatively on human wellbeing through agricultural, economical losses, and food security (Möller and Stukenbrock, 2017). Despite that, thanks to the development of the High-throughput sequencing (HTS), we are entering in a new era in plant pathology in which whole-genome sequences, or at least a reduced representation of the genome, of many individuals of pathogen species, are becoming readily available (Grünwald *et al.*, 2016). As a consequence, the scientific research questions that can be addressed, especially for non-model organisms, are being changed and today it became possible to take new challenges that were inconceivable thus far. However, this new reality brought some challenges, particularly in data analysis, and today, the bottleneck lies on the translation of this breadth of data into biological insights. Especially, due to the lack of gold standards or even an agreement as to which are the best strategies for processing and interpreting massive data sets (Pavey *et al.*, 2012).

Even though, the increase of data available due to the genomics revolution, allows the change of significant conceptual breakthroughs in how we view genetics, evolution, and the emergence of plant pathogens (Grünwald *et al.*, 2016). In fact, I believe that evolutionary biology can and should have a crucial role on the development of sustainable agricultural measures. Therefore, a combination of evolutionary analysis of genome-wide patterns, such as the demography history and evolutionary potential of a pathogen, joined together with the evaluation of which are the genomic regions associated with fungal pathogenicity and aggressiveness, will bring to light crucial information to guide the design of novel and sustainable strategies to slow down the emergence and spread of pathogens.

It was in this context, that the current thesis was planned and developed, trying to use the data generated by HTS to better understand the evolutionary features of *C. kahawae* and supply useful information for future sustainable disease control measures in *C. arabica*.

6.2 Conclusions and main contributions

6.2.1 Demographic history

In chapter 2 of this thesis, we investigated the demographic history and evolutionary potential of *C. kahawae,* using a genome-wide approach. When this work was initiated, the demographic history of *C. kahawae* had already been studied using a population genetic approach, with only a few loci and based on 3 SNPs. Despite the low genetic variation reported in such study, three completely differentiated populations were described (Angolan, Cameroonian and East African), from which Angolan population was described as the cradle of *C. kahawae's* origin followed by its sequential spreading to Cameroon and East Africa (Silva *et al.*, 2012). Moreover, it was also suggested that *C. kahawae* could be a true clonal pathogen perfectly adapted to green coffee berries (Silva *et al.*, 2012). However, the low genetic variability observed raised the question if a genomic approach would be able to detect a different demographic pattern.

To tackle this issue, we performed a demographic study using a population genome-wide approach with RADseq. During this work, unexpected results were obtained, such as the existence of two clonal lineages within Angolan population that gave rise to the Cameroonian population. The evolutionary mechanism responsible for this scenario is not evident, leading to the proposal of alternative hypotheses. Unfortunately, the lack of a reference genome did not allow the confirmation or rejection of such hypotheses, and this issue needs to be confirmed after the release of a reference genome with an appropriate annotation and chromosomal location. Besides that, it has been confirmed that only three completely differentiated populations are observed within *C. kahawae*. The most probable colonization scenario suggests that Angolan and East African populations emerged, virtually at the same time, and Cameroonian population further emerged from the Angolan population. Probably, the Angolan population remains as the

cradle of CBD, however the hypothesis that *C. kahawae* could have emerged from an unknown location and subsequently disseminated, almost simultaneously, to the Angola and East African plantations is now raised. Therefore, a subsequent sampling effort in Angola, especially regarding the ancestral lineages, would be required to completely unveil this mystery. Finally, it has been shown that *C. kahawae* is a true clonal pathogen, perfectly adapted to green coffee berries, which has a low evolutionary potential and dispersal ability, being the human transport its greatest dispersion factor. Thus, if quarantine measures are applied successfully, the potential dispersion of this harmful pathogen to other continents out-side Africa can be prevented. Additionally, due to the low evolutionary potential of this harmful pathogen a significant effort should be made on the development of *C. arabica* resistance varieties as well as on the implementation of sustainable agronomic measures able to reduce the disease incidence, such as good aeration, shade control, artificial irrigation during the drying season, removal of mummified berries and inter-planted coffee trees with fruits trees. Hopefully this combined approach may be enough to reduce the disease incidence without stimulating *C. kahawae* evolutionary mechanisms to overcome plant defenses, and thereby reduce the disease impact in coffee production. To conclude, this work also reinforces the importance of genomic studies, comprising a large number of loci, to capture a more accurate demographic history and the true evolutionary potential of a pathogen, since results based on a smaller number of loci can be affected by the lineage sorting history of particular loci, leading to a biased interpretation of results.

6.2.2 Aggressiveness profiling

In chapter 3 of this thesis, the focus was shifted for one of the most interesting phenotypic traits of *C. kahawae*, the aggressiveness. The aggressiveness was the first trait that allowed a complete differentiation of several isolates within the *C. kahawae* species (Beynon *et al.*, 1995; Loureiro *et al.*, 2011; Manga *et al.*, 1997; Pires *et al.*, 2016; Várzea *et al*., 1999), but the measures and plant material used to evaluate this trait could change according to the study. In fact until now, no comprehensive characterization of *C. kahawae's* aggressiveness, using a large set of isolates and metrics, has been performed and this kind of study is of the utmost importance not only to define a set of metrics, ranges and conditions able to accurately classify the isolates

aggressiveness, but also to create the necessary conditions for developing subsequent studies on the genetic mechanisms underlying this trait, such as genome-wide association studies.

It was then, with this focus, that the present work was carried out, being its main contributions: i) the release of a set of metrics and aggressiveness classes able to characterize other isolates, whether they are preserved in collections or recently collected from the field; ii) confirm the suitability of hypocotyls and detached green berries to perform *C. kahawae* aggressiveness assays, pointing out that hypocotyls are a more reproducible testing material then green berries; iii) perform the current classification of several *C. kahawae* isolates present on CIFC's collection, and with that bring the opportunity to perform association studies with higher depth between molecular markers and aggressiveness; and iv) show that aggressiveness is related with the development of post-penetration stages, rather than conidia germination and appressoria differentiation.

The importance of this work is unquestionable and without it, it would not have been possible to perform a genome-wide association study for aggressiveness, but its true impact relies on supporting coffee breeding programs. The selection of tester isolates representative of the global *C. kahawae* range of aggressiveness will allow the improvement of pre-screening resistance tests within breeding programs for producing more effective coffee resistant varieties. Finally, it would be very interesting to test the set of established metrics in controlled conditions in the field. To conclude, it is well known that aggressiveness is a very sensitive trait that can change according to the physiological state of the pathogen, inoculation, temperature, host physiology and plant material, and therefore our results should not be interpreted as an unvarying aggressiveness profile characterization, but rather as a current classification in a controlled environment.

6.2.3 Genomics of adaptation

In chapter 4, we investigated the genomic regions underlying the pathogenicity and aggressiveness of *C. kahawae*. When this work began, it was already known that *C. kahawae* had probably emerged through host-jump from a closely related non

pathogenic group. The genetic similarity between *C. kahawae* and this group was so high, that for some genes, the genetic information between them is completely shared.

For this reason, Weir *et al*., (2012) described this new group as a *C. kahawae* subspecies (*C. kahawae* subsp. c*iggaro*), based only in the genealogical concordance of phylogenetic species recognition criteria (Taylor *et al.*, 2000). However, in this work, this group is considered as a distinct taxa, more particularly, a cryptic species (*C. ciggaro*), in accordance with Batista *et al.*, (2017). Since, pathologically, they could not be more different, with *C. kahawae* being a specialist pathogen perfectly adapted to green coffee berries and restricted to Africa, while *C. ciggaro* is a generalist pathogen able to infect different hosts across the world (Silva *et al.*, 2012). On the other hand, the *C. kahawae'* aggressiveness variation, was already known (Beynon *et al.,* 1995; Loureiro *et al.*, 2011; Manga *et al.*, 1997; Pires *et al.*, 2016; Várzea *et al.*, 1999), and some studies had suggested that this trait could be under-selection (Boedo *et al.*, 2012; Pariaud *et al.,* 2009).

Therefore, *C. kahawae* was considered, in our study, as a perfect model to unveil the genomic bases underlying pathogenicity and aggressiveness patterns. Nonetheless, the identification of these genomic regions could be impaired due to *C. kahawae' s* true clonal lifestyle (without recombination) and the severe bottleneck effect that it suffered on its origin. In such a scenario, in which the demographic signals can mask the selection pattern, we attempted a possible solution, comparing pathogenic and non-pathogenic fungi, to look for the loci that have an excess of functional changes. This strategy, although not ideal, revealed some of the biological processes putatively involved in the pathogenicity of *C. kahawae* to *C. arabica*, even without a properly annotated reference genome. Of which I would like to highlight several interesting findings that were completely unknown until now: i) the genetic differentiation between pathogenic and non-pathogenic fungi is extremely high, reinforcing the distinction of *C. kahawae* as a separated species; ii) signatures of potential positive selection were found and those loci was retrieved as pathogenicity - related candidate genes; iii) the majority of the candidate genes are associated with transporters, oxidative response, and signaling; iv) 15% of the genes potentially under selection were described as having an important role in fungal pathogenicity and virulence, being some of them

identified as genes responsible for the total loss of pathogenicity in other fungi; v) the high abundance of transcription factors suggests that changes in gene expression patterns may be more important than the presence/absence of individual gene alleles; vi) the host specificity in closely related pathosystems of the *Colletotrichum* genus can be not only a matter of recognition, involving in particular pathogenicity factors for attempting penetration, but also related to a much broader adaptation to the living host environment across the entire course of pathogen development.

For the genome-wide association study with the phenotypic trait of aggressiveness no causal SNPs were detected, nonetheless, the two strategies applied allowed the identification of 10 SNPs and 15 SNPs of small effect in single and multi-association analysis, respectively, from which 7 were common. The annotation of the genomic regions containing these SNPs provided several candidate genes (F-box domain containing, nitrosoguanidine resistance, Fungal specific transcription factor domain-containing and C6 transcription factor) that can be putatively associated with aggressiveness. The presence of two transcription factors in this list may suggest that gene expression has, in fact, an important role on aggressiveness regulation. Moreover, the detection of only SNPs of small effect could mean that: i) aggressiveness is regulated by a set of small effect SNPs that are very difficult to detect only with an association study and other approaches, such as transcriptomics, are further required; ii) aggressiveness is a plastic trait regulated by gene expression and associated regulatory mechanisms; iii) this trait is not under-selection and is governed only by physiological conditions. Despite that, as referred above, the clonal nature of *C. kahawae* poses some limitations to this kind of analyses, and therefore our results should be interpreted with caution. Even though, we have now, a list of candidate genes that can be studied through gene expression and additional functional analyses (knockouts, knockdowns and transgenics) to ascertain their role in *C. kahawae* aggressiveness and pathogenicity. Finally, it is important to note, that this is the first work that attempted to understand this complex interaction from the pathogen's point of view, and this knowledge could be useful to the development of sustainable control measures.

6.2.4 Follow-up studies through gene expression

In chapter 5 of this thesis, we established the best normalization strategy to perform gene expression studies in *C. kahawae* using a wide range of fungal and interaction samples and aggressiveness profiles. At the time that this work started, it was already known the best normalization strategy to perform gene expressions studies through qPCR in *C. kahawae – C. arabica* interaction, but only from the plant perspective (Figueiredo *et al.*, 2013). However, this thesis aimed to validate the functional role of the candidate genes and compare its expression pattern between isolates with different aggressivenesses profiles, and consequently, it was necessary to choose the most appropriate normalization strategy to perform such analyses. In this study, the candidate reference genes were chosen based on two strategies, a genome wide approach (RNA-seq datasets) and the literature, and became evident that RNA-seq data proved to be a useful alternative source of reference genes. No global normalization strategy could be established for all tested conditions, but rather two different sets of reference genes are needed to normalize *C. arabica - C. kahawae* interaction samples and *C. kahawae* samples, respectively. Altogether, this study provided, for the first time, the tools required to conduct accurate qPCR studies in *C. kahawae* considering its aggressiveness patterns, developmental stages and host interactions. At this point, it is possible to study the set of candidate genes, identified in our work, that are putatively related to pathogenicity and associated with aggressiveness, by assessing its differential expression profiles.

6.3 Future perspectives

Overall, the results of this thesis demonstrate the power of applying an integrative approach, focused on evolutionary biology, to investigate and better understand host-pathogen interaction processes, the population dynamics and evolutionary potential of a pathogen. CBD is still a poorly studied disease and an effort should be made to improve sustainable control measures as well as to increase the knowledge on *C. kahawae* pathogenicity and aggressiveness. In this sense, this thesis should be seen as the first steps to better understand the genomics underlying these biological processes, and an additional sequencing effort is required. It would be of note the genome sequencing of

C. kahawae and *C. ciggaro* to perform a comparative genomic analysis, and a transcriptomic analysis of isolates with distinct aggressiveness profiles in several crucial phases of the infection process. The combination of both approaches will help to understand the pathogenicity and aggressiveness of this harmful pathogen. In fact, with these resources, I believe that *C. kahawae* will become a wonderful model to study some very interesting evolutionary mechanisms, including host-jump, the evolutionary dead end of a true clonal pathogen, and the role of selection on aggressiveness. Finally, the most exciting outcomes of this thesis are the novel questions and challenges that came to light, especially: i) the footprints of ancient episodes of introgression, which can suggest that in some unfavorable conditions *C. kahawae* is able to exchange genetic information between clonal lineages, and therefore adapt more quickly to a new environment. ii) test, through gene expression and additional functional analyses, the list of candidate genes putatively associated with the pathogenicity and aggressiveness of *C. kahawae*; iv) re-sequencing of genome and epigenome to better understand the biological mechanisms underlying the aggressiveness of *C. kahawae*; iii) the importance of clarifying the taxonomic classification of *C. kahawae* reinforcing its distinction as a separated species.

6.4 References

Batista, D., Silva, D.N., Vieira, A., *et al*. (2017) Legitimacy and implications of reducing *Colletotrichum kahawae* to subspecies in plant pathology. *Front. Plant Sci.* **7**, 1–4.

Beynon, S.M., Coddington, B. and Varzea, V. (1995) Genetic variation in the coffee berry disease pathogen, *Colletotrichum kahawae*. *Physiol. Mol. Plant Pathol.* **46**, 457–470.

Boedo, C., Benichou, S., Berruyer, R., *et al*. (2012) Evaluating aggressiveness and host range of *Alternaria dauci* in a controlled environment. *Plant Pathol.* **61**, 63–75.

Figueiredo, A., Loureiro, A., Batista, D., Monteiro, F., Várzea, V., Pais, M.S., Gichuru, E.K. and Silva, M.C. (2013) Validation of reference genes for normalization of qPCR gene expression data from *Coffea* spp. hypocotyls inoculated with *Colletotrichum kahawae*. *BMC Res. Notes* **6**, 388.

Grünwald, N.J., McDonald, B.A.M. and Milgroom, M.G.M.G. (2016) Population genomics of fungal and Oomycete pathogens. *Annu. Rev. Phytopathol.* **54**, 323–346.

Loureiro, A., Guerra-Guimarães, L., Lidon, F.C., Bertrand, B., Silva, M.C. and Várzea, V. (2011) Isoenzymatic characterization of *Colletotrichum kahawae* isolates with different levels of aggressiveness. *Trop. Plant Pathol.* **36**, 287–293.

Manga, M., Bieysse, D., Mouen-Bedimo, J., Akalay, I., Bompard, E. and Berry, D. (1997) Observation sur la diversité de la population de *Colletotrichum kahawae* agent de l'anthracnose des baies du cafeier Arabica. Implications pour l'am? eloration getique. *Proc. 17th Int. Conf. Coffee Sci. 1997, Nairobi, Kenya. Paris, Fr. Assoc. Sci. Int. du Cafe*, 604–12.

McDonald, B.A. and Stukenbrock, E.H. (2016) Rapid emergence of pathogens in agro-ecosystems: global threats to agricultural sustainability and food security. *Philos. Trans. R. Soc. B Biol. Sci.* **371**, 20160026.

Möller, M. and Stukenbrock, E.H. (2017) Evolution and genome architecture in fungal plant pathogens. *Nat. Rev. Microbiol.* **15**, 756–771.

Pariaud, B., Ravigné, V., Halkett, F., Goyeau, H., Carlier, J. and Lannou, C. (2009) Aggressiveness and its role in the adaptation of plant pathogens. *Plant Pathol.* **58**, 409–424.

Pavey, S.A., Bernatchez, L., Aubin-Horth, N. and Landry, C.R. (2012) What is needed for next-generation ecological and evolutionary genomics? *Trends Ecol. Evol.* **27**, 673–676.

Pires, A.S., Azinheira, H.G., Cabral, A., *et al*. (2016) Cytogenomic characterization of *Colletotrichum kahawae*, the causal agent of coffee berry disease, reveals diversity in minichromosome profiles and genome size expansion. *Plant Pathol.* **65**, 968–977.

Silva, D.N., Talhinhas, P., Cai, L., Manuel, L., Gichuru, E.K., Loureiro, A., Várzea, V., Paulo, O.S. and Batista, D. (2012) Host-jump drives rapid and recent ecological speciation of the emergent fungal pathogen *Colletotrichum kahawae*. *Mol. Ecol.* **21**, 2655–2670.

Taylor, J.W., Jacobson, D.J., Kroken, S., Kasuga, T., Geiser, D.M., Hibbett, D.S. and Fisher, M.C. (2000) Phylogenetic species recognition and species concepts in fungi. *Fungal Genet. Biol.* **31**, 21–32.

Várzea, V., Rodrigues, C.J., Silva, M., Pedro, J. and Marques, D. (1999) High virulence of a *Colletotrichum kahawae* isolate from Cameroon as plant pathology compared with other isolates from other regions. *InProceedings 18th Int. Conf. Coffee Sci. 1999, Helsinki, Finland. Paris, Fr. Assoc. Sci. Int. du Cafe*, 131.

Weir, B.S., Johnston, P.R. and Damm, U. (2012) The *Colletotrichum gloeosporioides* species complex. *Stud. Mycol.* **73**, 115–180.

Appendix

1

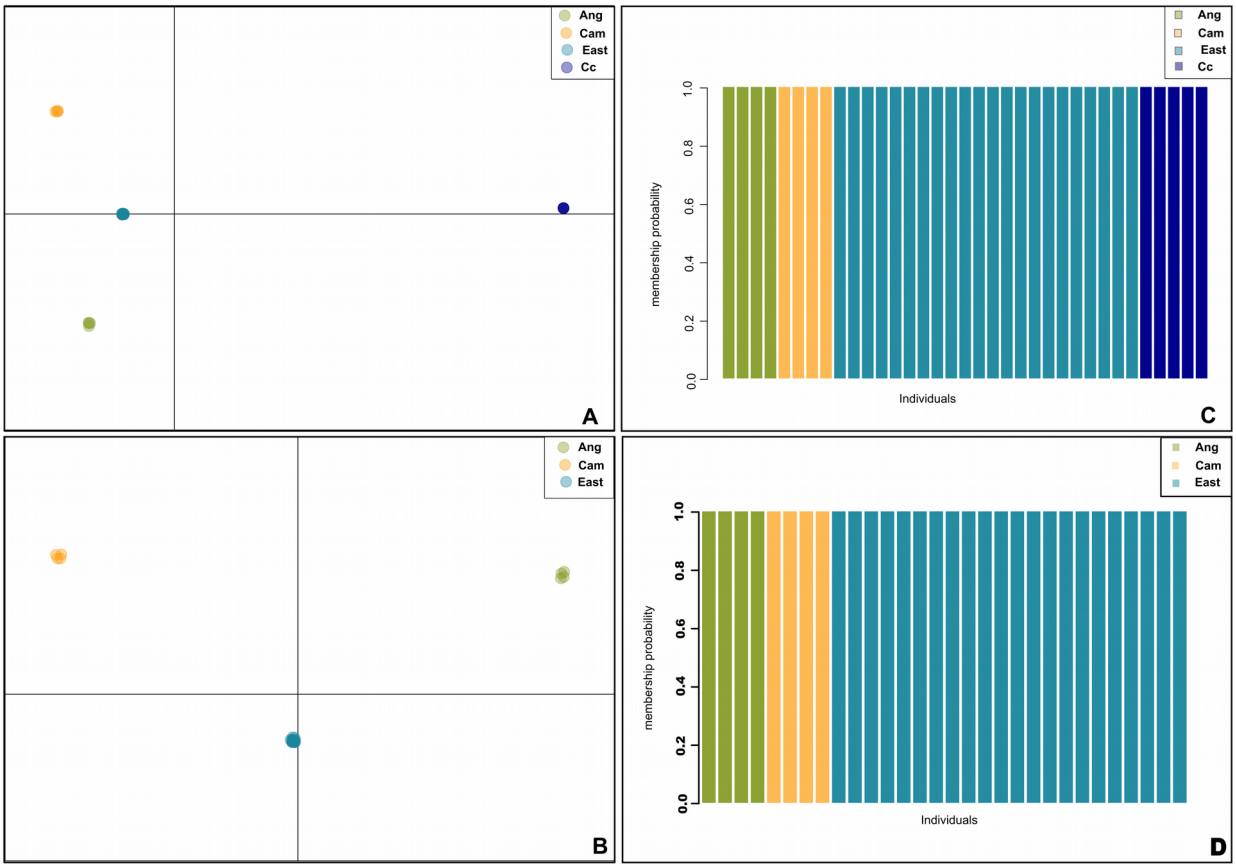
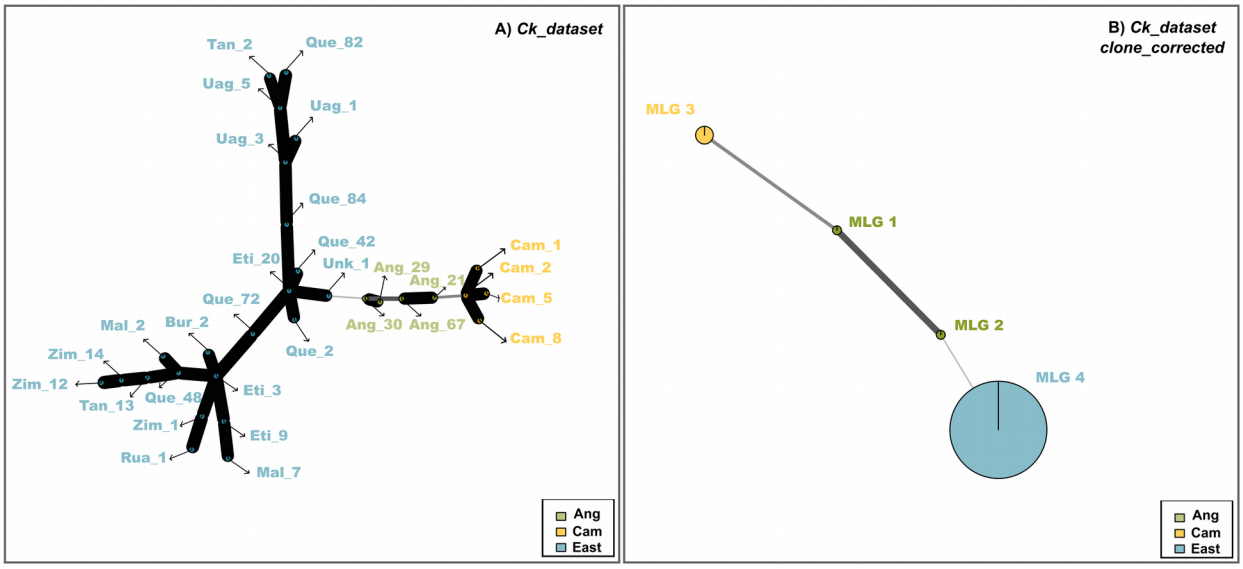


Figure A1.1 – Scatterplot of the DAPC analyses using the *ck_dataset* and *total_dataset.* A) Scatterplot of the discriminant analysis of principal components for total_dataset. B) Scatterplot of the discriminant analysis of principal components for ck_dataset. For A and B, only the two-first principal components of the DAPC are represented. The first axis is the horizontal axis. C) Clustering of 35 isolates representing world-wide geographical distribution of *C. kawahae* and *C. ciggaro.* D) Clustering of 30 isolates representing world-wide geographical distribution of *C. kawahae*. The colours represent the groups found by the k-means methods and legend is provided in the figure. [Angolan (Ang), Cameroonian (Cam) and East African (East)]



igure A1.2 - Minimum spanning networks of *C. kawahae*. **A)** *ck_dataset*. **B)** *ck_clone_corrected_dataset*. When the dataset was clone-corrected, only four clonal lineages or (MLG) were detected: two in Angola, one in Cameroon and one in East Africa. The colors represent the three *C. kawahae* populations according to the legend is provided. [Angolan (Ang), Cameroonian (Cam) and East African (East)]

Table A1.1 - List of the isolates with information regarding their species, natural host, geographic origin and year of collection

Isolates	Species/Group	Host	Country/Region	Collected year
Ang_21	*C. kahawae*	*C. arabica*	Angola/Amboim	2004
Ang_29	*C. kahawae*	*C. arabica*	Angola/Ganda	2005
Ang_30	*C. kahawae*	*C. arabica*	Angola/Ganda	2005
Ang_67	*C. kahawae*	*C. arabica*	Angola/Ganda	2005
Bur_2	*C. kahawae*	*C. arabica*	Burundi/NAa	1992
Cam_1	*C. kahawae*	*C. arabica*	Cameroon/Babadjou	1992
Cam_2	*C. kahawae*	*C. arabica*	Cameroon/Santa	1992
Cam_5	*C. kahawae*	*C. arabica*	Cameroon/Baham	1996
Cam_8	*C. kahawae*	*C. arabica*	Cameroon/Kumbo	1996
Eti_9	*C. kahawae*	*C. arabica*	Ethiopia/Sidamo	1993
Eti_20	*C. kahawae*	*C. arabica*	Ethiopia/NA	1993
Eti_3	*C. kahawae*	*C. arabica*	Ethiopia/NA	1993
Mal_2	*C. kahawae*	*C. arabica*	Malawi/NA	1988
Mal_7	*C. kahawae*	*C. arabica*	Malawi/NA	1993
Que_2	*C. kahawae*	*C. arabica*	Kenya/NA	1989
Que_48	*C. kahawae*	*C. arabica*	Kenya/Taita Taveta	1996
Que_72	*C. kahawae*	*C. arabica*	Kenya/Ruiru	2001
Que_82	*C. kahawae*	*C. arabica*	Kenya/Kitale	2010
Que_84	*C. kahawae*	*C. arabica*	Kenya/Mgumguri	2010
Que_42	*C. kahawae*	*C. arabica*	Kenya/NA	1996
Rua_1	*C. kahawae*	*C. arabica*	Rwanda/Gicumbo	1989
Tan_2	*C. kahawae*	*C. arabica*	Tanzania/Ngoro	2006
Tan_13	*C. kahawae*	*C. arabica*	Tanzania/Mbinga	2006
Uga_1	*C. kahawae*	*C. arabica*	Uganda/Kapchorwa	2010
Uga_3	*C. kahawae*	*C. arabica*	Uganda/Kapchorwa	2010
Uga_5	*C. kahawae*	*C. arabica*	Uganda/Kapchorwa	2010
Zim_12	*C. kahawae*	*C. arabica*	Zimbabwe/NA	1997
Zim_1	*C. kahawae*	*C. arabica*	Zimbabwe/Hiton	1991
Zim_14	*C. kahawae*	*C. arabica*	Zimbabwe/NA	1997
Unk_1	*C. kahawae*	*C. arabica*	East African/NA	1992
Cc_1275.8 (ICMP 17922) [a]	*C. ciggaro* (*C. kahawae* subsp ciggaro*)	*Hypericum perforatum*	Germany/NA	1937
Cc_1206.3 (ICMP 12953)[a]	*C. ciggaro* (*C. kahawae* subsp *ciggaro**)	*Persea americana*	New Zealand/NA	1991
Cc_1252.12 (ICMP 18534)[a]	*C. ciggaro* (*C. kahawae* subsp *ciggaro**)	*Kunzea ericoides*	New Zealand/NA	x
Cc_1262.12 (ICMP 18539)[a]	*C. ciggaro* (*C. kahawae* subsp *ciggaro**)	*Olea europaea*	Australia/NA	1989
Cc_432[b]	*C. ciggaro* (*C. kahawae* subsp *ciggaro**)	*Mangifera indica*	Portugal/Lisbon	2010

C. kahawae and *C. ciggaro* are accepted as two cryptic species as suggested by Batista *et al.* (2017).

* as characterized in Weir *et al.*, 2016

[a]were kindly provided by Bevan Weir and Peter Johnston (Landcare Research, Auckland, New Zealand), and [b] by Ana Paula Ramos (Instituto Superior de Agronomia, Universidade de Lisboa, Lisboa, Portugal).

Table A1.2 – Fst values obtained with a pairwise analysis between the three populations of *C. kawahae* [Angolan (Ang), Cameroonian (Cam) and East African (East)]and between the two cryptic species

	Ang	**Cam**	**East**
Ang			
Cam	0.80715		
East	0.9603	0.98188	

	Cc
Ck	0.88652

Table A1.3 – Mapping and annotation of the segregated SNPs alleles within the two Angola clonal lineages

Clone_a

A – Segregated SNPs of clone_a

Locus name	Genomic region	Type of mutation	Scaffold	Gene position	Segregated SNP	Description
25378	exon	non-synonymous	scaffold141	36767	G/C	hypothetical protein
47573	exon	non-synonymous	scaffold334	56357	T/C	caib baif family enzyme
28823	exon	non-synonymous	scaffold259	54479	G/T	hec ndc80p family protein
25183	exon	non-synonymous	scaffold38	2008	C/T	amidohydrolase
369	exon	non-synonymous	scaffold851	48806	G/C	FAD dependent oxidoreductase superfamily protein
42905	exon	non-synonymous	scaffold176	5847	G/A	serin endopeptidase
4959	exon	non-synonymous	scaffold653	1013	C/T	chitin synthase activator
33417	exon	non-synonymous	scaffold334	86973	T/A	coatomer zeta subunit
5334	exon	non-synonymous	scaffold236	299712	C/T	hypothetical protein
46067	exon	non-synonymous	scaffold508	19246	G/A	hypothetical protein
3213	exon	non-synonymous	scaffold143	29573	G/A	pre-mRNA splicing factor ATP-dependent RNA helicase prp43
12741	exon	non-synonymous	scaffold64	16110	G/A	hypothetical protein
27542	exon	non-synonymous	scaffold27	29627	G/A	hypothetical protein
45510	exon	non-synonymous	scaffold875	36913	C/T	efflux pump antibiotic resistance
15857	exon	synonymous	scaffold236	179656	C/T	heterokaryon incompatibility protein
17149	exon	synonymous	scaffold64	3357	C/T	zinc fyve domain containing protein
42424	exon	synonymous	scaffold401	27831	G/A	beta-ig-h3 fasciclin
22788	exon	synonymous	scaffold962	12819	C/T	fes cip4 domain-containing protein
24519	exon	synonymous	scaffold236	131204	G/A	autophagy protein
11427	exon	synonymous	scaffold494	45748	C/T	hypothetical protein
38742	exon	synonymous	scaffold236	276278	G/A	ras gtpase activating protein
34034	exon	synonymous	scaffold364	270690	G/C	bar domain-containing protein
42950	exon	synonymous	scaffold25	82689	C/T	cbf nf-y family transcription factor
31037	exon	x	scaffold236	213849	C/T	iq calmodulin-binding motif protein
3453	exon	x	scaffold565	89294	G/T	MFS sugar transporter
42784	exon	x	scaffold55	18415	G/A	glutamyl-tRNA amidotransferase
45763	exon	x	scaffold62	78400	G/A	alpha mannosyltransferase
31529	intergenic	x	scaffold589	26958	A/C	C6 zinc finger domain-containing protein'])

B - SNPs derived from the ancestral lineage and shared between the Clone_a and Cameroon

Locus name	Genomic region	Type of mutation	Scaffold	Gene position	Segregated SNP	Description
44569	exon	non-synonymous	scaffold104	35225	G/A	ubiquitin-like activating enzyme

37724	exon	non-synonymous	scaffold850	61549	A/C	plasma membrane channel protein
25213	exon	synonymous	scaffold25	84791	G/C	pre-mRNA-splicing factor 38a
50738	exon	synonymous	scaffold72	28965	A/C	spo76 protein
13701	exon	synonymous	scaffold154	124362	G/A	protein efr3
46765	intergenic	x	scaffold102	32492	G/A	x
14462	intergenic	x	scaffold219	34421	C/T	x
317	intergenic	x	scaffold411	29247	C/T	x
1402	intergenic	x	scaffold153	84843	G/A	x
23717	intergenic	x	scaffold169	53855	A/C	x
8664	intergenic	x	scaffold323	17910	G/C	x
1516	intergenic	x	scaffold574	63392	G/A	x

Clone_b

C – Segregated SNPs of clone_b

Locus name	Genomic region	Type of mutation	Scaffold	Gene position	Segregated SNP	Description
6087	exon	non-synonymous	scaffold304	141522	G/A	ribosomal protein l24e
15757	exon	non-synonymous	scaffold14	51187	C/T	urea amidolyase
270	exon	non-synonymous	scaffold412	143168	C/T	major facilitator superfamily transporter
39406	exon	non-synonymous	scaffold718	60477	C/T	duf1212 domain membrane protein
2269	exon	non-synonymous	scaffold460	44353	G/T	pro1 protein
28357	exon	non-synonymous	scaffold599	28088	G/C	FAD dependent oxidoreductase %2C putative
24422	exon	non-synonymous	scaffold292	82016	G/C	sugar transport protein 4
18574	exon	non-synonymous	scaffold358	117941	G/A	glycoside hydrolase family 43
9133	exon	non-synonymous	scaffold95	85513	C/T	f420-dependent nadp reductase
20123	exon	non-synonymous	scaffold114	21295	A/C	putative plant-like oligopeptide transporter
49947	exon	non-synonymous	scaffold863	3894	C/T	MFS drug transporter
43732	exon	non-synonymous	scaffold524	31028	G/T	telomere length regulation protein
34874	exon	non-synonymous	scaffold154	171142	G/A	DNA repair protein
41286	exon	non-synonymous	scaffold478	65221	G/A	trichothecene efflux pump
20097	exon	non-synonymous	scaffold522	63503	C/T	TPR domain-containing protein
14909	exon	non-synonymous	scaffold568	54808	G/A	extracellular dihydrogeodin oxidase laccase
52281	exon	non-synonymous	scaffold506	29797	C/T	telomere length regulator protein
17462	exon	non-synonymous	scaffold402	91340	C/T	p-type ATPase
46050	exon	non-synonymous	scaffold667	68379	C/T	efflux pump antibiotic resistance
27411	exon	synonymous	scaffold150	84590	G/C	bZIP transcription factor
35780	exon	synonymous	scaffold266	171664	T/A	urease accessory protein
48242	exon	synonymous	scaffold673	133503	G/A	glycosyl hydrolase
52115	exon	synonymous	scaffold237	80056	G/A	succinate dehydrogenase flavoprotein subunit

31339	exon	synonymous	scaffold213	150811	G/A	3-beta hydroxysteroid dehydrogenase isomerase family
32915	exon	synonymous	scaffold304	165786	C/G	extracellular serine-threonine rich protein
30232	exon	synonymous	scaffold282	96422	A/C	mitochondrial chaperone bcs1
37993	intergenic	x	scaffold565	80293	A/C	drug resistance protein
2802	intergenic	x	scaffold574	33675	C/T	enoyl-hydratase'])
25902	intergenic	x	scaffold58	159171	G/A	methyltransferase domain-containing protein'])
39907	intergenic	x	scaffold60	116558	A/C	MFS multidrug transporter'])
45789	intergenic	x	scaffold66	29048	G/A	GNAT family'])

D - SNPs derived from the ancestral lineage and shared between the Clone_b and Cameroon

Locus name	Genomic region	Type of mutation	Scaffold	Gene position	Segregated SNP	Description
32954	exon	non-synonymous	scaffold98	149135	G/A	hypothetical protein
18992	exon	non-synonymous	scaffold43	82616	G/A	beta-ig-h3 fasciclin
50050	exon	non-synonymous	scaffold581	15355	C/T	yap-binding protein
26873	exon	synonymous	scaffold395	16767	C/T	kinesin related protein 2
28038	exon	synonymous	scaffold364	275726	C/T	beta-ig-h3 fasciclin
37369	gene	x	scaffold646	32120	G/A	hypothetical protein
17841	intergenic	x	scaffold399	146204	G/A	x
26273	intergenic	x	scaffold88	42717	G/A	x
30985	intergenic	x	scaffold385	13750	A/T	x
47930	intergenic	x	scaffold112	51337	A/C	x
1245	intergenic	x	scaffold385	9729	C/T	x
6834	intergenic	x	scaffold116	18447	C/T	x
7821	intergenic	x	scaffold754	40015	G/A	x
10064	intergenic	x	scaffold754	31217	C/G	x

Appendix

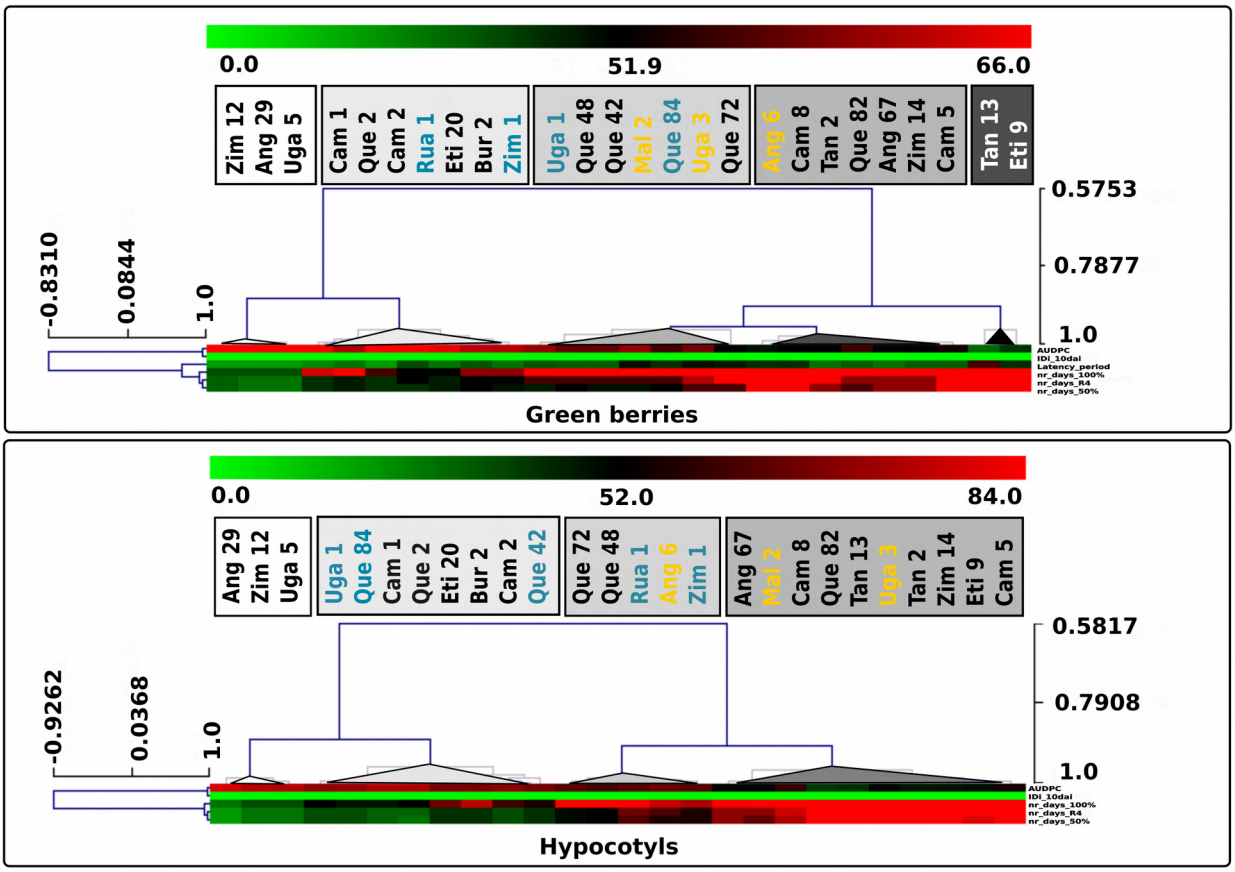


Figure A2.1- *C. kahawae* isolate group clustering from a heatmap analysis, using the data from all quantitative traits recorded in green berries (a) and hypocotyls (b). Isolate groups are presented in colour coded boxes corresponding to different aggressiveness classes (high - white; high_moderate - light grey; low_moderate - grey; low - dark gray). Isolates whose aggressiveness classification varies according to the data collected in green berries or hypocotyls are color highlighted (blue - changes only between the two moderate sub-classes; yellow - changes between the three main classes)

Table A2.1 - Correlation coefficient analysis between all the aggressiveness quantitative traits recorded in hypocotyls (light grey) and detached green berries (dark grey), based on the mean value of the two assays for each isolate

	RAUDPC	AUDPC	IDI_10dai	nr_days_R4	nr_days_50%	nr_days_100 %	Latent_perio d	Incubation_perio d	
RAUDPC		r= 1.00; p= <0.00001	r= 0.98; p= <0.00001	r= -0.93; p= <0.00001	r=-0.95; p= <0.00001	r= -0.90; p= <0.00001	-	r= -0.26; p=0.192157	**Hypocotyls**
AUDPC	r= 1.00; p= <0.00001		r= 0.96; p= <0.00001	r=-0.95; p= <0.00001	r=-0.96 p= <0.00001	r= -0.91; p= <0.00001	-	r= -0,26; p=0.191808	
IDI_10dai	r= 0.94; p= <0.00001	r= 0.94; p= <0.00001		r= -0.96 p= <0.00001	r= -0.92; p= <0.00001	r= -0.91; p= <0.00001	-	r= -0.25; p=0.212392	
nr_days_R4	r= -0.96; p= <0.00001	r= -0.97; p= <0.00001	r= -0.92; p= <0.00001		r= 0.99 p= <0.00001	r= 0.93; p= <0.00001	-	r= 0.19; p=0.357725	
nr_days_50%	r= -0.96; p= <0.00001	r= -0.96; p= <0.00001	r= -0.91; p= <0.00001	r= 0.95; p= <0.00001		r= 0.92; p= <0.00001	-	r= 0.20; p=0.325561	
nr_days_100%	r= -0.90; p= <0.00001	r= -0.91; p= <0.00001	r= -0.91; p= <0.00001	r=0.92; p= <0.00001	r=0.91; p= <0.00001		-	r=0.30; p=0.130813	
Latent_period	r= -0.72; p= 3.8E-05	r= -0.73; p= 2.3E-05	r= -0.75; p= 1.2E-05	r=0.79 p= <0.00001	r= 0.72; p= 3.3E-05	r= 0.67; p= 0.000179		-	
Incubation_perio d	r= -0.70 p= 7.9E-05	r= -0.70 p= 7.9E-05	r= -0.65 p= 0.000349	r=0.64; p= 0.000409	r=0.65; p= 0.000285	r=0.58 p= 0.001958	r=0.42 p=0.033		
	green berries								

Table A2.2- Detailed data description of all aggressiveness quantitative traits recorded in hypocotyls and detached green berries for each *C. kahawae* isolate, and final assignment to aggressiveness classes

Isolates	Aggressiveness quantitative traits															Class	
	RAUDPC		AUDPC		IDI_10dai		nr_days_R4		nr_days_50%		nr_days_100%		Latent_period	Incubation_period			
	green berries	hypocotyls	green berries	hypocotyls	green berries	hypocotyls	green berries	hypocotyls	green berries	hypocotyls	green berries	hypocotyls	green berries	green berries	hypocotyls	green berries	hypocotyls
Ang 29	18.27 ± 0.65	19.86 ± 0.07	74.71 ± 0.33	79.42 ± 0.28	1 ± 0	1 ± 0	8 ± 00	6 ± 0	8 ± 0	6 ± 0	10 ± 0	8 ± 0	6 ± 0	3 ± 0	3 ± 0	high	high
Zim 12	18.49 ± 0.06	18.92 ± 0.65	73.95 ± 0.23	75.69 ± 2.60	1 ± 0	1 ± 0	9 ± 1.41	8 ± 0	9 ± 1.41	8 ± 00	10 ± 0	10 ± 0	6 ± 0	3 ± 0	3 ± 0	high	high
Uga 5	18.33 ± 0.09	18.80 ± 0.10	73.31 ± 0.36	75.20 ± 0.42	1 ± 0	1 ± 0	8 ± 00	8 ± 0	8 ± 0	8 ± 00	10 ± 0	10 ± 0	6 ± 0	3 ± 0	3 ± 0	high	high
Uga 1	15.07 ± 2.91	17.57 ± 0.31	60.29 ± 11.64	70.30 ± 1.22	0.65 ± 0.30	0.91 ± 0.04	15 ± 2.83	10 ± 0	13 ± 3.54	10 ± 0	20 ± 0	14 ± 1.41	10 ± 4.95	3 ± 0	6 ± 0	low_moderate	high_moderate
Que 84	13.95 ± 2.14	17.87 ± 0.16	55.80 ± 8.56	71.47 ± 0.63	0.47 ± 0.25	0.93 ± 0	15 ± 2.83	10 ± 0	14 ± 1.41	10 ± 0	20 ± 3.54	13 ± 0	12 ± 2.12	6 ± 0	3 ± 0	low_moderate	high_moderate
Cam 1	16.41 ± 1.54	18.72 ± 0.15	65.64 ± 6.16	74.86 ± 0.61	0.77 ± 0.22	0.97 ± 0	12 ± 2.12	9 ± 1.41	12 ± 2.12	9 ± 1.41	17 ± 0	13 ± 0	7 ± 1.41	5 ± 2.12	3 ± 0	high_moderate	high_moderate
Que 2	15.43 ± 2.59	18.48 ± 0.52	61.72 ± 10.37	73.92 ± 2.07	0.67 ± 0.38	0.96 ± 0.04	13 ± 3.54	9 ± 1.41	13 ± 3.54	8 ± 00	18 ± 3.54	14 ± 1.41	8 ± 2.83	6 ± 0	6 ± 0	high_moderate	high_moderate
Eti 20	16.25 ± 2.02	16.95 ± 1.00	65.01 ± 8.08	67.81 ± 4.02	0.77 ± 0.21	0.85 ± 0.09	13 ± 3.54	12 ± 2.12	13 ± 3.54	12 ± 2.12	14 ± 1.41	18 ± 3.54	8 ± 0	5 ± 2.12	6 ± 0	high_moderate	high_moderate
Cam 2	16.74 ± 0.22	17.61 ± 0.21	66.98 ± 0.89	70.45 ± 0.84	0.88 ± 0.06	0.91 ± 0	12 ± 2.12	10 ± 0	12 ± 2.12	10 ± 0	15 ± 00	16 ± 1.41	6 ± 0	3 ± 0	5 ± 2.12	high_moderate	high_moderate
Que 42	14.35 ± 1.00	16.69 ± 1.51	57.40 ± 4.00	66.78 ± 6.05	0.50 ± 0.04	0.82 ± 0.19	15 ± 00	12 ± 2.12	13 ± 0	12 ± 2.12	19 ± 2.12	15 ± 2.85	9 ± 1.41	3 ± 0	5 ± 2.12	low_moderate	high_moderate
Bur 2	15.65 ± 1.12	16.96 ± 1.40	62.61 ± 4.47	67.84 ± 5.60	0.64 ± 0.17	0.85 ± 0.12	14 ± 1.41	12 ± 2.12	14 ± 1.41	12 ± 2.12	16 ± 1.41	20 ± 0	10 ± 4.95	6 ± 0	6 ± 0	high_moderate	high_moderate
Rua 1	16.49 ± 0.66	15.71 ± 1.89	65.97 ± 2.64	62.83 ± 7.56	0.76 ± 0.08	0.73 ± 0.15	14 ± 1.41	18 ± 6.36	12 ± 2.12	17 ± 4.95	14 ± 1.41	(22 - >24) ± a	10 ± 0	3 ± 0	6 ± 0	high_moderate	low_moderate
Que 72	13.86 ± 1.27	15.78 ± 2.36	55.45 ± 5.09	63.17 ± 9.40	0.50 ± 0.18	0.76 ± 0.28	16 ± 1.41	14 ± 4.95	15 ± 0	14 ± 4.95	21 ± 1.41	22 ± 2.83	9 ± 1.41	6 ± 0	7 ± 1.41	low_moderate	low_moderate
Ang 6	12.02 ± 1.38	16.32 ± 0.55	48.08 ± 5.54	65.27 ± 2.18	0.34 ± 0.15	0.83 ± 0.05	19 ± 2.12	17 ± 4.95	18 ± 3.54	15 ± 7.07	21 ± 1.41	21 ± 4.95	12 ± 2.12	6 ± 0	5 ± 2.12	low	low_moderate
Que 48	14.38 ± 0.33	15.12 ± 1.22	57.51 ± 1.33	60.50 ± 4.86	0.56 ± 0.06	0.69 ± 0.14	15 ± 00	14 ± 1.41	14 ± 1.41	14 ± 1.41	20 ± 0	22 ± 2.83	9 ± 1.41	6 ± 0	6 ± 0	low_moderate	low_moderate
Zim 1	15.53 ± 0.57	15.73 ± 1.90	62.10 ± 2.27	62.94 ± 7.61	0.66 ± 0.12	0.80 ± 0.18	14 ± 1.41	15 ± 7.07	13 ± 0	15 ± 7.07	15 ± 2.83	19 ± 7.78	8 ± 0	6 ± 0	7 ± 1.41	high_moderate	low_moderate

Que 82	12.91 ± 3.96	9.78 ± 1.86	51.63 ± 15.84	40.54 ± 9.44	0.35 ± 0.29	0.42 ± 0.20	18 ± 3.54	>24 ± –	15 ± 2.85	>24 ± 0	19 ± 4.95	>24 ± 0	13 ± 3.54	6 ± 3.54	6 ± 2.12	**low**	**low**
Tan12	12.97 ± 2.77	11.82 ± 1.43	51.88 ± 11.07	47.27 ± 5.71	0.48 ± 0.25	0.41 ± 0.24	18 ± 3.54	(20 - >24) ± a	16 ± 1.41	22 ± 2.83	(22 - >24) ± a	>24 ± 0	8 ± 0	6 ± 0	8 ± 2.83	**low**	**low**
Tan 13	7.61 ± 3.12	10.43 ± 0.02	30.44 ± 12.48	41.71 ± 0.07	0.26 ± 0.09	0.40 ± 0.15	(24 - >24) ± a	(24 - >24) ± a	(20 - >24) ± a	(24 - >24) ± a	>24 ± 0	>24 ± 0	15 ± 7.07	7 ± 1.41	8 ± 2.83	**very_ low**	**low**
Mal 2	15.09 ± 0.19	12.94 ± 1.58	60.36 ± 0.75	51.76 ± 6.34	0.63 ± 0.06	0.44 ± 0.18	15 ± 00	19 ± 2.12	13 ± 0	17 ± 0	19 ± 2.12	24 ± 0	13 ± 0	6 ± 0	6 ± 0	**low_moderat e**	**low**
Ang 67	12.09 ± 1.39	12.89 ± 1.23	51.03 ± 9.30	51.54 ± 4.94	0.39 ± 0.19	0.48 ± 0.03	16 ± 1.41	19 ± 2.12	16 ± 1.41	19 ± 2.12	20 ± 3.54	24 ± 0	8 ± 2.83	6 ± 0	6 ± 0	**low**	**low**
Eti 9	10.26 ± 3.78	12.27 ± 0.34	41.03 ± 15.13	49.09 ± 1.34	0.30 ± 0.19	0.49 ± 0.07	(20 - >24) ± a	22 ± 0	18 ± 3.54	21 ± 1.41	>24 ± 0	>24 ± 0	12 ± 2.12	7 ± 1.41	3 ± 0	**very_ low**	**low**
Uga 3	14.64 ± 0.83	12.65 ± 1.03	58.58 ± 3.33	50.62 ± 4.13	0.51 ± 0.02	0.54 ± 0.07	17 ± 4.95	>24 ± 0	15 ± 2.85	>24 ± 0	21 ± 1.41	>24 ± 0	13 ± 0	3 ± 0	5 ± 2.12	**low_moderat e**	**low**
Zim 14	13.03 ± 0.90	9.30 ± 1.28	52.13 ± 3.59	47.65 ± 5.62	0.41 ± 0.09	0.52 ± 0.01	16 ± 1.41	(20 - >24) ± a	16 ± 1.41	(20 - >24) ± a	21 ± 1.41	>24 ± 0	9 ± 1.41	6 ± 0	6 ± 0	**low**	**low**
Cam 8	13 ± 2.13	14.62 ± 0.98	52.09 ± 8.50	58.46 ± 3.91	0.39 ± 0.09	0.64 ± 0.08	18 ± 3.54	(15 - >24) ± a	18 ± 3.54	(13 - >24) ± a	(22 - >24) ± a	>24 ± 0	12 ± 2.12	3 ± 0	5 ± 2.12	**low**	**low**
Cam 5	13.27 ± 2.10	13.66 ± 1.82	53.07 ± 8.14	54.64 ± 7.27	0.50 ± 0.16	0.63 ± 0.22	18 ± 3.54	22 ± 2.83	18 ± 3.54	22 ± 2.83	(20 - >24) ± a	>24 ± 0	9 ± 1.41	6 ± 0	5 ± 2.12	**low**	**low**

a* not computed

Table A2.3 - Correlation coefficient analysis of pairwise comparisons between experimental assays for each *C. kawahae* isolate, considering both hypocotyls and detached green berries, independently

Isolates	green berries		hypocotyls	
	Pearson Correlation Coefficient (r)	p-value	Pearson Correlation Coefficient (r)	p-value
Ang29_(E1*E2)	0.95	2.3E-05	1.00	< 0.00001
Zim 12_(E1*E2)	0.99	< 0.00001	0.99	< 0.00001
Uga 5_(E1*E2)	0.99	< 0.00001	0.98	< 0.00001
Uga 1_(E1*E2)	0.88	0.000853	0.99	< 0.00001
Que 84_(E1*E2)	0.96	< 0.00001	0.97	< 0.00001
Cam 1_(E1*E2)	0.94	6.5E-05	0.99	< 0.00001
Que 2_(E1*E2)	0.85	0.001647	0.99	< 0.00001
Eti 20_(E1*E2)	0.93	7.7E-05	0.98	< 0.00001
Cam 2_(E1*E2)	0.99	< 0.00001	1.00	< 0.00001
Que 42_(E1*E2)	0.99	< 0.00001	0.92	0.000188
Bur 2_(E1*E2)	0.97	< 0.00001	0.98	< 0.00001
Rua 1_(E1*E2)	0.98	< 0.00001	0.98	< 0.00001
Que 72_(E1*E2)	0.97	< 0.00001	0.88	0.000678
Ang 6_(E1*E2)	0.98	< 0.00001	0.96	< 0.00001
Que 48_(E1*E2)	0.97	< 0.00001	0.97	< 0.00001
Zim 1_(E1*E2)	0.98	< 0.00001	0.90	0.000421
Que 82_(E1*E2)	0.91	0.000278	0.88	0.000888
Tan 12_(E1*E2)	0.94	4.9E-05	0.92	0.000195
Tan 13_(E1*E2)	0.97	< 0.00001	0.94	5.2E-05
Mal 2_(E1*E2)	0.99	< 0.00001	0.94	5.3E-05
Ang 67_(E1*E2)	0.85	0.001652	0.98	< 0.00001
Eti 9_(E1*E2)	0.95	2.6E-05	0.98	< 0.00001
Uga 3_(E1*E2)	0.96	< 0.00001	0.93	8.8E-05
Zim 14_(E1*E2)	0.98	< 0.00001	0.89	0.00051
Cam 8_(E1*E2)	0.98	< 0.00001	0.96	< 0.00001
Cam 5_(E1*E2)	0.98	< 0.00001	0.85	0.001926

Table A2.4 - Correlation coefficient analysis within and between experimental assays of hypocotyls and detached green berries based on AUDPC values

	Pearson Correlation Coefficient (r)	p-value
hypocotyls_E1* hypocotyls_E2	0.82	< 0.00001
green berries_E1* green berries_E2	0.59	0.001511
hypocotyls_average (E1andE2) * green berries_average (E1andE2)	0.77	< 0.00001

Table A2.5 - Non-parametric Mann Whitney test, at significance level of 1%, on *AUDPC* values comparing the three main established aggressiveness classes *(high, moderate, low*) and sub-classes (*high_moderate and low_moderate*)

Mann Whitney (MW)	p-values
low*high	0.0001
moderate*high	0.0001
moderate*low	0.0001
low_moderate* high	0.0001
high_moderate*high	0.0001
high_moderate*low	0.0001
low_moderate*low	0.0002
high_moderate*low_moderate	0.0136

Appendix

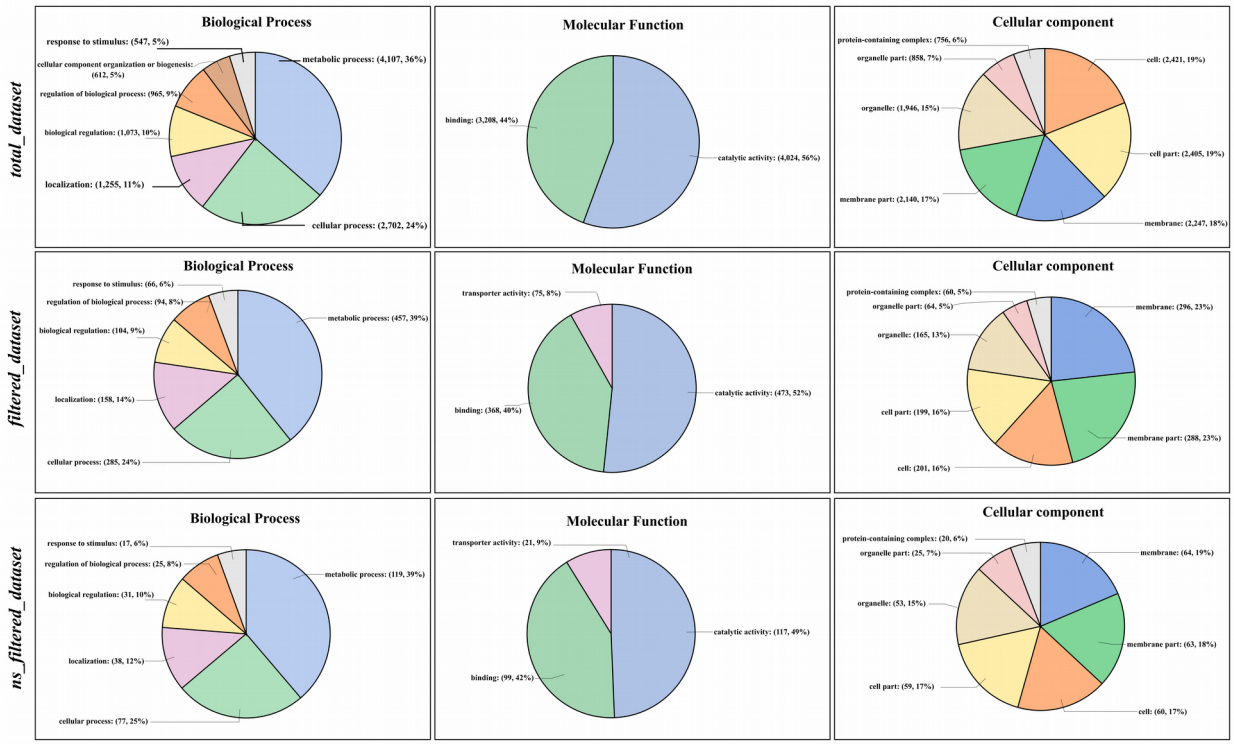


Figure A3.1 – Functional annotation level 2 comparative graph of Biological processes, Molecular Function and Cellular component to all datasets under study (*total_dataset*, *filtered_dataset* and *ns_filtered_dataset*)

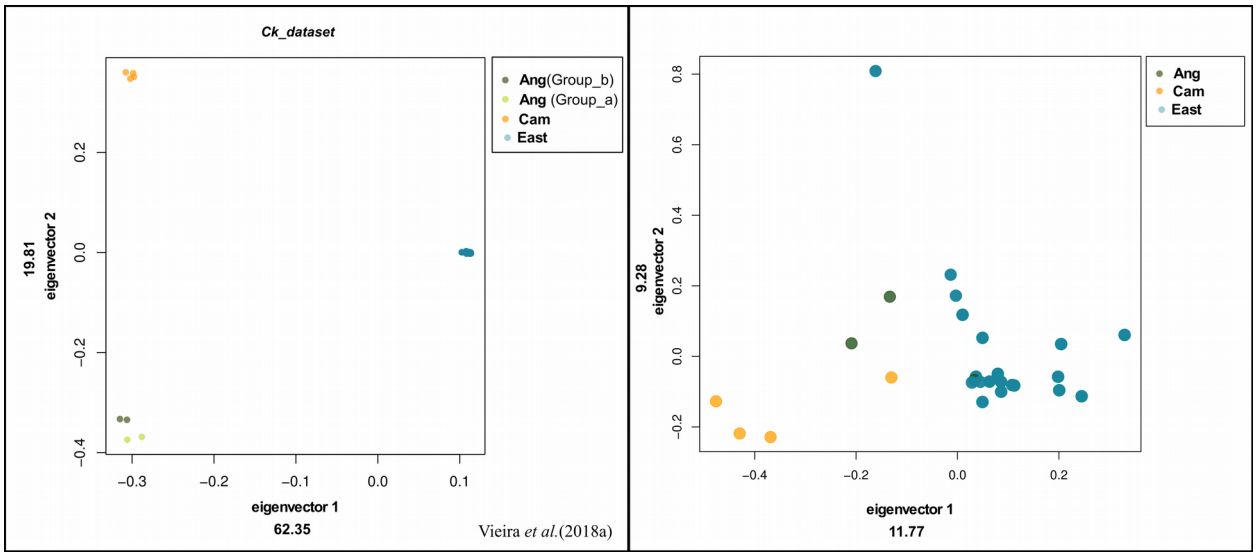


Figure A3.2 – Principal component analysis of genomic diversity within *C. kahawae*. a) all detected SNPs within *C. kahawae,* adapted from Chapter 2; b) *filtered_dataset*. The percentage of variation explained by each principal component is provided in their respective label. Isolates are color coded according to the respective population as provided in the legend

Table A3.1 - List of the isolates with information regarding their species, natural host, geographic origin and year of collection

Isolates	Species/Group	Host	Country/Region	Collected year
Ang21	*C. kahawae*	*C. arabica*	Angola/Amboim	2004
Ang29	*C. kahawae*	*C. arabica*	Angola/Ganda	2005
Ang30	*C. kahawae*	*C. arabica*	Angola/Ganda	2005
Ang67	*C. kahawae*	*C. arabica*	Angola/Ganda	2005
Bur2	*C. kahawae*	*C. arabica*	Burundi/NAa	1992
Cam1	*C. kahawae*	*C. arabica*	Cameroon/Babadjou	1992
Cam2	*C. kahawae*	*C. arabica*	Cameroon/Santa	1992
Cam5	*C. kahawae*	*C. arabica*	Cameroon/Baham	1996
Cam8	*C. kahawae*	*C. arabica*	Cameroon/Kumbo	1996
Eti9	*C. kahawae*	*C. arabica*	Ethiopia/Sidamo	1993
Eti20	*C. kahawae*	*C. arabica*	Ethiopia/NA	1993
Eti3	*C. kahawae*	*C. arabica*	Ethiopia/NA	1993
Mal2	*C. kahawae*	*C. arabica*	Malawi/NA	1988
Mal7	*C. kahawae*	*C. arabica*	Malawi/NA	1993
Que2	*C. kahawae*	*C. arabica*	Kenya/NA	1989
Que48	*C. kahawae*	*C. arabica*	Kenya/Taita Taveta	1996
Que72	*C. kahawae*	*C. arabica*	Kenya/Ruiru	2001
Que82	*C. kahawae*	*C. arabica*	Kenya/Kitale	2010
Que84	*C. kahawae*	*C. arabica*	Kenya/Mgumguri	2010
Que42	*C. kahawae*	*C. arabica*	Kenya/NA	1996
Rua1	*C. kahawae*	*C. arabica*	Rwanda/Gicumbo	1989
Tan2	*C. kahawae*	*C. arabica*	Tanzania/Ngoro	2006
Tan13	*C. kahawae*	*C. arabica*	Tanzania/Mbinga	2006
Uga1	*C. kahawae*	*C. arabica*	Uganda/Kapchorwa	2010
Uga3	*C. kahawae*	*C. arabica*	Uganda/Kapchorwa	2010
Uga5	*C. kahawae*	*C. arabica*	Uganda/Kapchorwa	2010
Zim12	*C. kahawae*	*C. arabica*	Zimbabwe/NA	1997
Zim1	*C. kahawae*	*C. arabica*	Zimbabwe/Hiton	1991
Zim14	*C. kahawae*	*C. arabica*	Zimbabwe/NA	1997
Unk1	*C. kahawae*	*C. arabica*	East Africa/NA	1992
C1275.8 (ICMP 17922) [a]	*C. ciggaro (C. kahawae subsp ciggaro*)*	*Hypericum perforatum*	Germany/NA	1937
C1206.3 (ICMP 12953)[a]	*C. ciggaro (C. kahawae subsp ciggaro*)*	*Persea americana*	New Zealand/NA	1991
C1252.12 (ICMP 18534)[a]	*C. ciggaro (C. kahawae subsp ciggaro*)*	*Kunzea ericoides*	New Zealand/NA	x
C1262.12 (ICMP 18539)[a]	*C. ciggaro (C. kahawae subsp ciggaro*)*	*Olea europaea*	Australia/NA	1989
Cg432b	*C. ciggaro (C. kahawae subsp ciggaro*)*	*Mangifera indica*	Portugal/Lisbon	2010
C880.1 (ICMP 18530)	*C. aotearoa*	*Vitex lucens*	New Zealand	1988
C1282.3 (ICMP 18536)	*C. aotearoa*	*Coprosma sp.*	New Zealand	2009
C1282.4 (ICMP 18537)	*C. aotearoa*	*Coprosma sp.*	New Zealand	2009
C1288.1 (ICMP 18541)	*C. aotearoa*	*Coprosma sp.*	New Zealand	2010
C1291 (ICMP 18542)	*Gomera cingulata* "f.sp.camelliae"	*Camellia sasanqua*	USA	2002

C. kahawae and *C. ciggaro* are accepted as two cryptic species as suggested by Batista *et al.* (2017).

* as described by Weir *et al.* (2012)

[a]were kindly provided by Bevan Weir and Peter Johnston (Landcare Research, Auckland, New Zealand), and [b] by Ana Paula Ramos (Instituto Superior de Agronomia, Universidade de Lisboa, Lisboa, Portugal).

Table A3.2 – Parameters tested for the *de novo* assembly and number of SNPs retrieved in each combination

Tested Parameters	Total SNPs	Total SNPs (MM100)	Total SNPs (MM80)	Total SNPs (MM50)
L8:5; L9:5; L10: 0.94; L12:5; L13:3	128607	29405	97456	107972
L8:10; L9:5; L10: 0.94; L12:5; L13:3	127978	29845	99228	107978
L8:10; L9:4; L10: 0.90; L12:5; L13:3	173911	43215	131903	143982
L8:20; L9:5; L10: 0.94; L12:5; L13:3	104521	6136	73017	89208
L8:5; L9:5; L10: 0.94; L12:40; L13:3	43000	41855	42956	42945
L8:10; L9:5; L10: 0.94; L12:40; L13:3	30723	29845	30699	30703
L8:10; L9:4; L10: 0.90; L12:40; L13:3	44795	43215	44703	44711
L8:5; L9:5; L10: 0.94; L12:5; L13:2	128607	29405	97456	107972
L8:10; L9:5; L10: 0.94; L12:5; L13:2	127978	29845	99228	107978
L8:10; L9:4; L10: 0.90; L12:5; L13:2	173911	43215	131903	143982

Table A3.3 - List of the genes potentially under selection, number of synonymous and non-synonymous mutations, dN/dS ratio, gene description and best hit on the blast analysis against the PHI-base

https://drive.google.com/open?id=178o4rJi1c-xOlsolxeZr5knGiAvOhquIzAkIDF0rsss

Table A3.4 - Detailed information on the candidate genes under selection with a hit on the PHI-base, identified as potentially involved in pathogenicity and virulence in other host-pathogen interactions

https://docs.google.com/spreadsheets/d/1bNdWMTJGo9l6fHUjglk5X7WH5Ui4qDpqzej_3Htehyw/edit?usp=sharing

Table A3.5 - Detailed information of the candidate genes putatively associated with aggressiveness with a hit on the PHI-base, identified as potentially involved in pathogenicity and virulence in other host-pathogen interactions

PHI-base								
Cf Gene ID	***Ck RAD loci***	**Gene_description**	**Pathogen gene (Cf Gene ID)**	**Pathogen gene (Ck RAD loci)**	**Pathogen Species**	**Disease**	**Host Species**	**Mutant Phenotype**
CGLO_08602	x	Fungal specific transcription factor domain-containing	FZC28	x	Magnaporthe oryzae	Rice blast	Oryza sativa (related: rice)	unaffected pathogenicity
x	35951		x	GzZC278	Fusarium graminearum	Fusarium ear blight	Triticum (related: wheat)	
CGLO_00627	x	C6 transcription factor	GzZC184		Fusarium graminearum	Fusarium ear blight	Triticum (related: wheat)	unaffected pathogenicity
	44503			FZC55	Magnaporthe oryzae	Rice blast	Oryza sativa (related: rice)	
CGLO_03059	x	hypothetical protein	SrbA	x	Aspergillus fumigatus	Pulmonary aspergillosis	Mus musculus (related: house mouse)	reduced virulence
x	17838		x	No hits	No hits	No hits	No hits	
x	x	---NA--	x	No hits	No hits	No hits	No hits	reduced virulence
x	46939		GIT3	No hits	Candida albicans	Disseminated candidiasis	Mus musculus (related: house mouse)	

Appendix

4

A4.1 – cDNA sequences for the genes under study.

https://doi.org/10.1371/journal.pone.0150651.s001

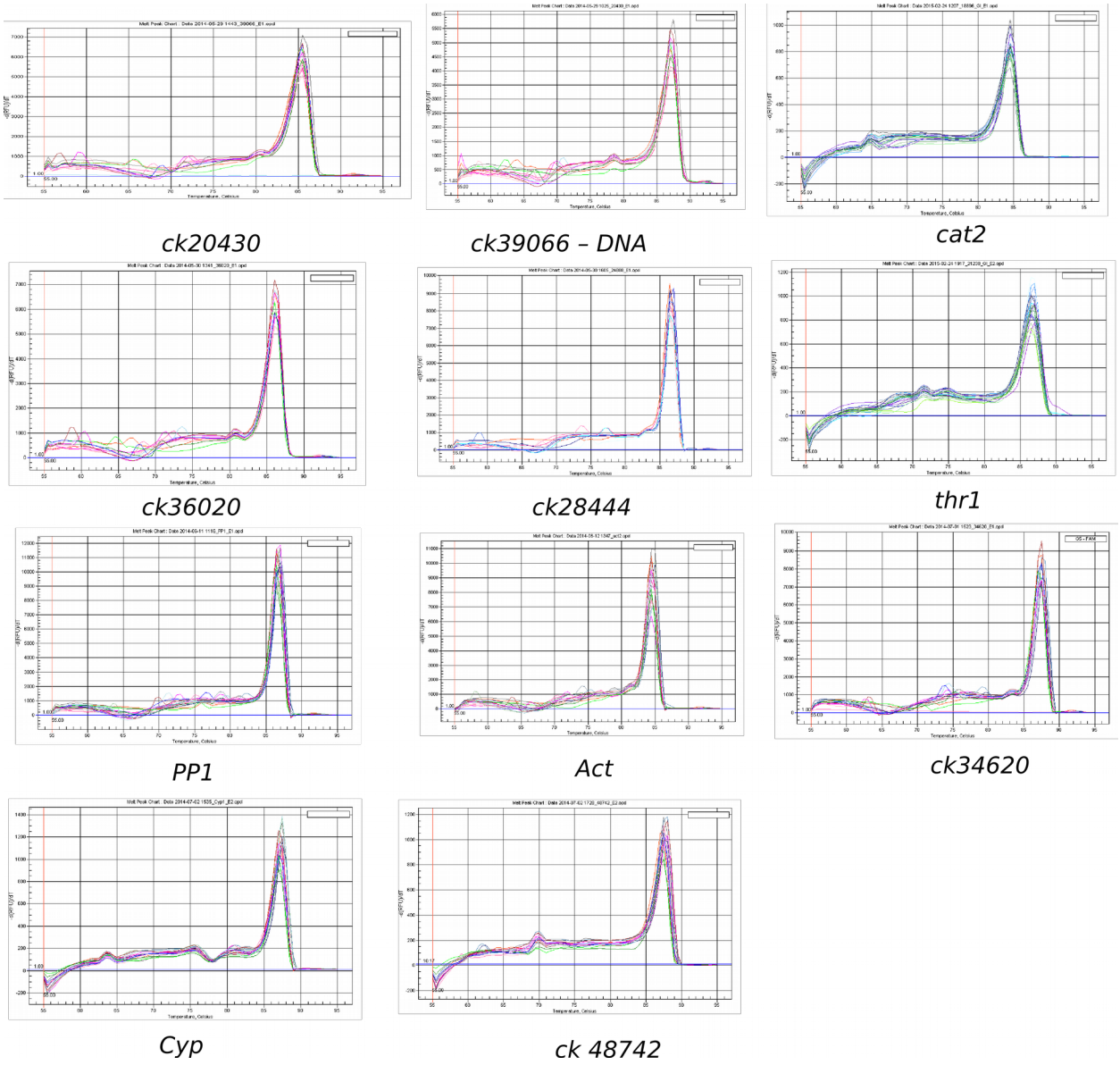


Figure A4.1 - Primer specificity test through dissociation curve analysis collected from iQ5 (Bio-rad) using several samples of *C. kahawae* and *C. arabica – C. kahawae.*

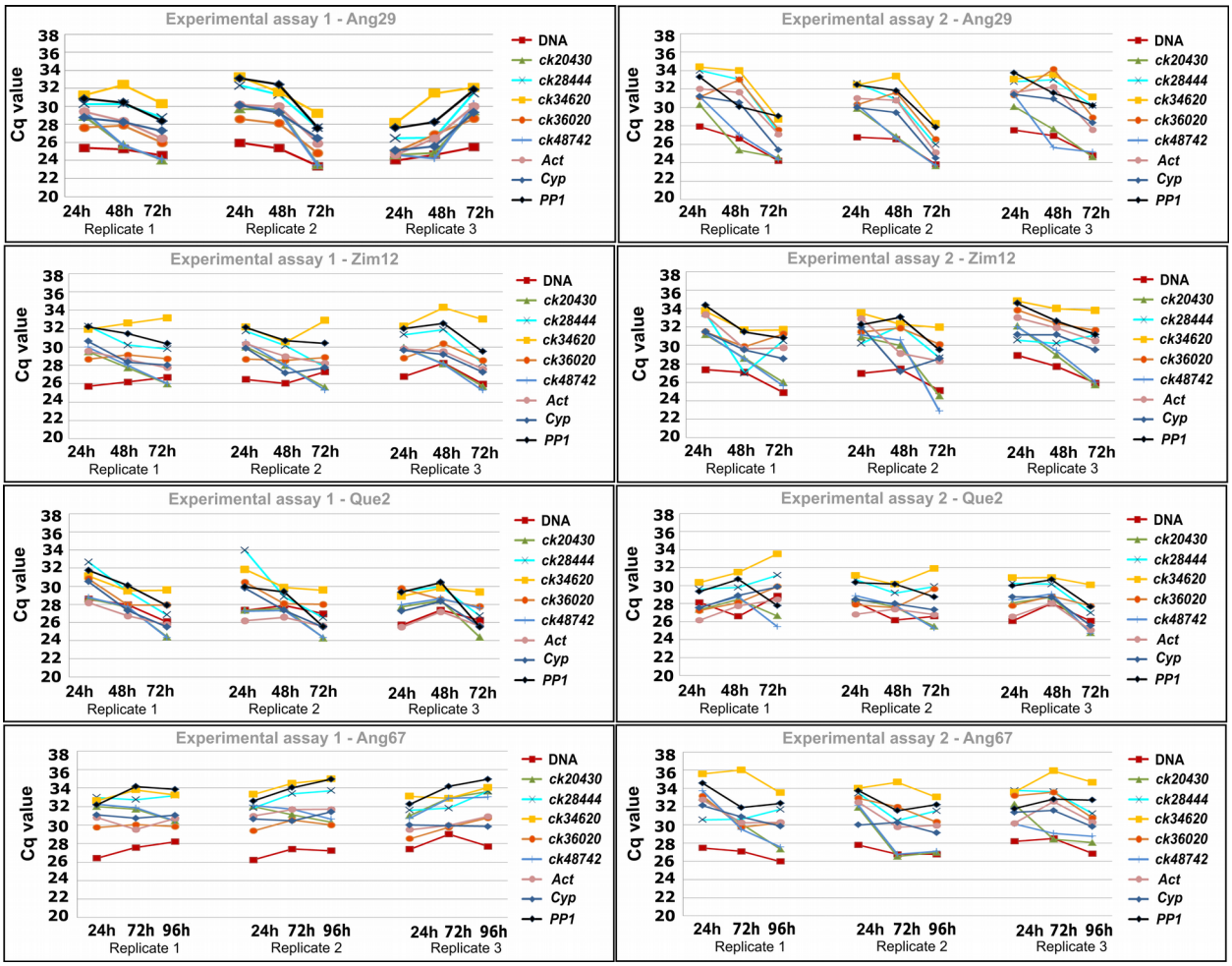


Figure A4.2 - Cq values of reference genes compared with fungal biomass normalization. RNA transcription levels of candidate reference genes tested during the infection time-course are presented as Cq mean value in the different samples, against the respective biomass quantification with Cq DNA value (ck39066), for two independent experiments.

Figure A4.3 - Relative quantification of *thr1* expression using six different normalization factors (NF). Expression profiles are presented per isolate (Ang29 (A), Zim 12 (B), Que2 (C) and Ang67 (D)), during the early stages of infection process and growth (Ap: Appressoria; M: Mycelium). Details on the normalization factors are described in table 4.

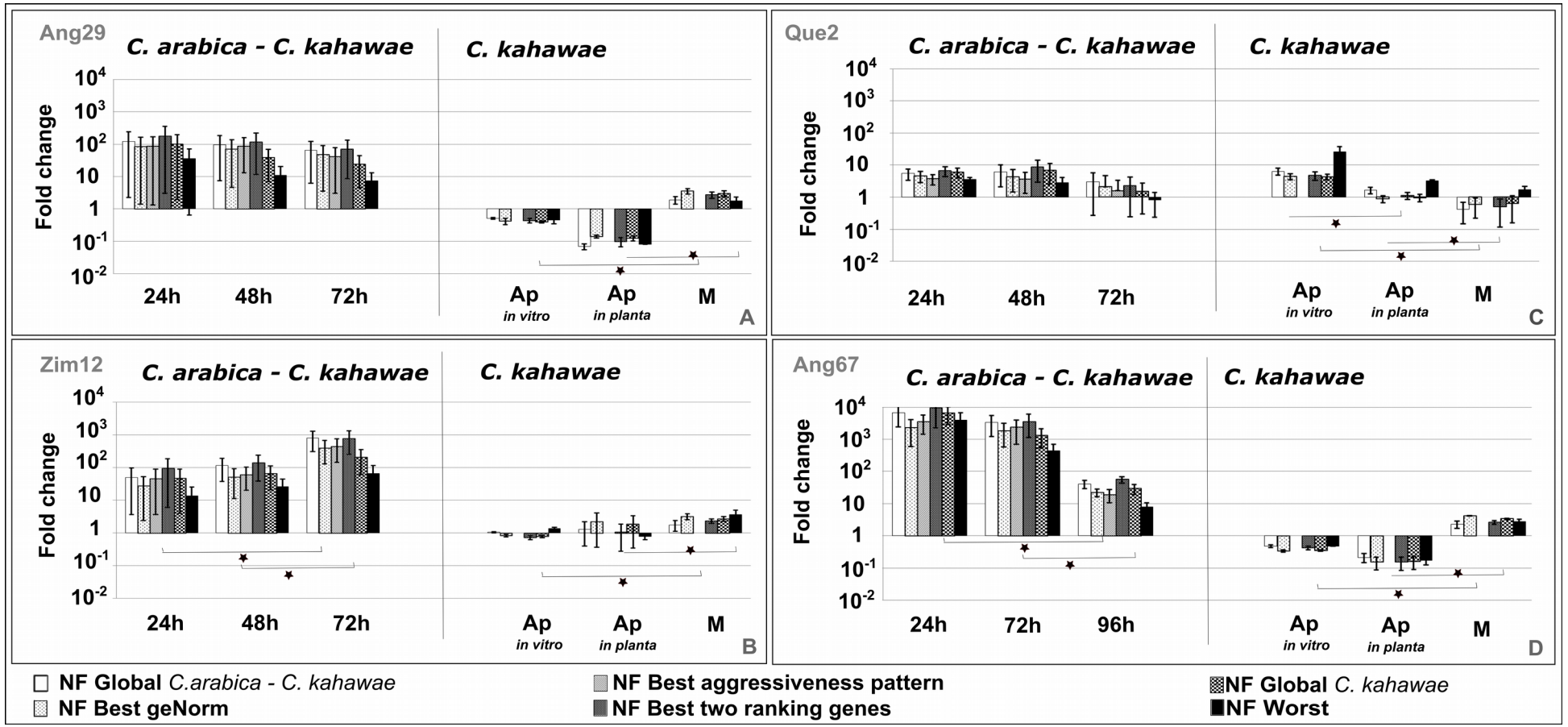


Figure A4.4 - Relative quantification of *cat2* expression using six different normalization factors (NF). Expression profiles are presented per isolate (Ang29 (A), Zim 12 (B), Que2 (C) and Ang67 (D)), during the early stages of infection process and growth (Ap: Appressoria; M: Mycelium). Details on the normalization factors are described in table 4.

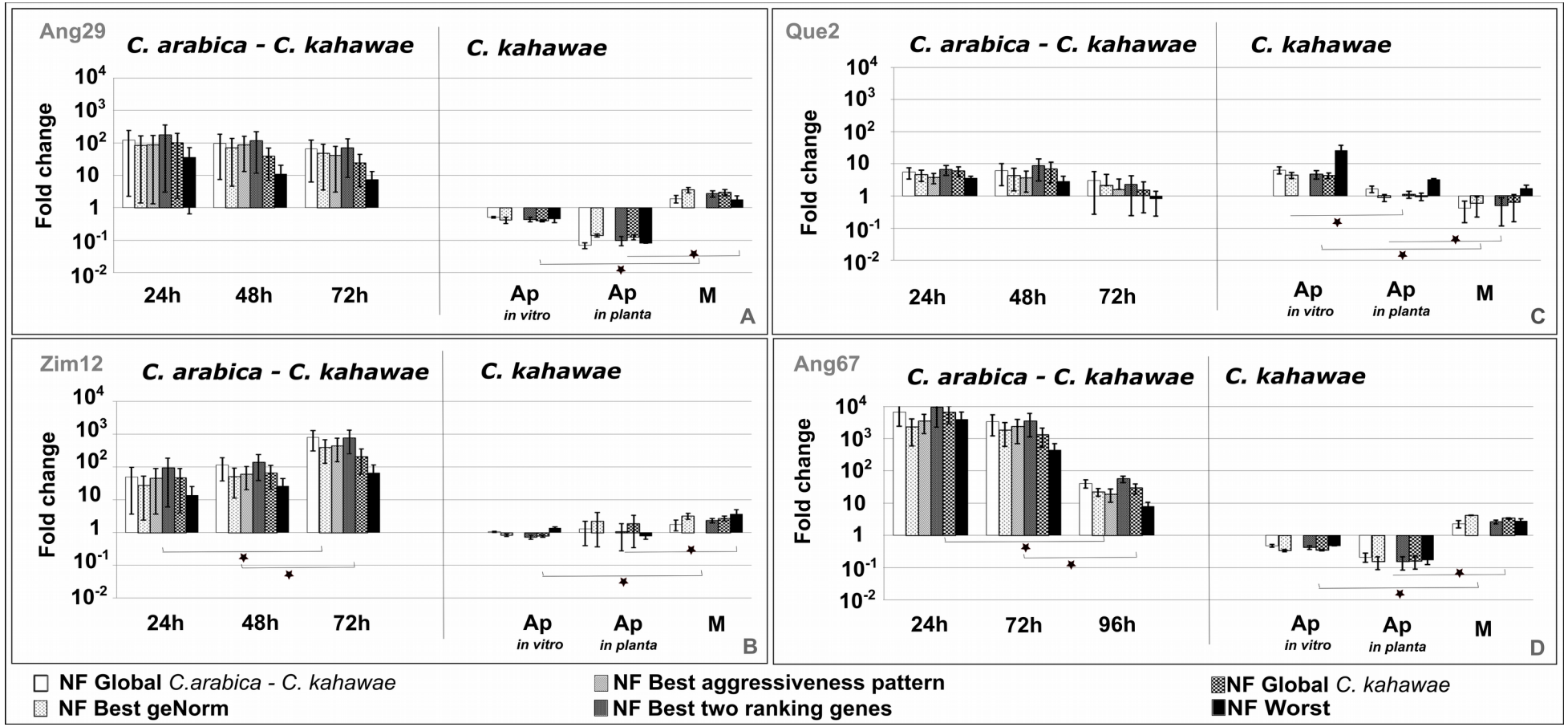


Table A4.1 – Primer efficiency specific for the type of samples under study (*C. arabica - C. kahawae* samples and *C. kahawae* samples)

			C. arabica - C.kahawae	*C.kahawae*	Average	SD
Reference genes	**Literature**	*Act*	1.918	1.886	1.902	0.023
		Cyp	1.916	1.907	1.910	0.006
		PP1	1.915	1.924	1.919	0.006
	RNA-seq	*ck20430*	1.916	1.930	1.923	0.010
		ck28444	1.957	1.908	1.933	0.034
		ck48742	1.946	1.911	1.929	0.025
		ck36020	1.919	1.941	1.930	0.016
		ck39066	1.923	1.894	1.907	0.021
		ck34620	1.934	1.921	1.926	0.009
Genes of interest	**RNA-seq**	*ck21238*	1.950	1.950	1.950	0.000
		ck25805	1.923	1.920	1.921	0.002

Table A4.2. - Ranking of the candidate reference genes for the *C. arabica - C. kahawae* samples according to the isolates under study. Stability values and ranking of candidate reference genes given by geNorm and Normfinder are provided alongside with the overall ranking calculated by the arithmetic mean ranking value of each gene using the two applets. Genes were ranked from the most stable (1) to the least stable (8).

Ang 29					
Gene Name	**NormFinder**		**geNorm**		**Overall ranking**
	Stability value	**Ranking**	**M value**	**Ranking**	
ck48742	0.76	8	1.37	6	8
ck20430	0.63	6	1.5	7	7
ck36020	0.71	7	1.17	5	6
ck28444	0.27	2	0.51	1	2
Cyp	0.22	1	0.51	1	1
Act	0.32	3	0.64	2	3
ck34620	0.57	5	1	4	5
PP1	0.33	4	0.85	3	4
Zim12					
Gene Name	**NormFinder**		**geNorm**		**Overall ranking**
	Stability value	**Ranking**	**M value**	**Ranking**	
ck48742	0.93	8	1.59	7	8
ck20430	0.74	6	1.51	6	7
ck36020	0.73	5	1.13	3	4
ck28444	0.6	4	1.34	5	5
Cyp	0.5	3	0.78	1	3
Act	0.39	2	0.78	1	1
ck34620	0.8	7	1.18	4	6
PP1	0.2	1	0.96	2	1
Que 2					
Gene Name	**NormFinder**		**geNorm**		**Overall ranking**
	Stability value	**Ranking**	**M value**	**Ranking**	
ck48742	0.6	6	1.28	7	6
ck20430	0.52	4	0.66	1	3

ck36020	0.67	7	1.23	6	6
ck28444	0.58	5	1.1	4	5
Cyp	0.25	1	0.91	2	1
Act	0.46	2	1.02	3	3
ck34620	0.73	8	1.18	5	6
PP1	0.49	3	0.66	1	2

Ang 67

Gene Name	NormFinder		geNorm		Overall ranking
	Stability value	Ranking	M value	Ranking	
ck48742	0.57	7	1.5	7	7
ck20430	0.66	8	1.36	6	7
ck36020	0.51	6	1.23	5	6
ck28444	0.42	5	1.1	4	5
Cyp	0.29	1	0.72	1	1
Act	0.3	2	0.83	2	2
ck34620	0.42	3	0.72	1	2
PP1	0.42	4	1	3	4

Table A4.3 - Statistical analysis of *thr1* expression in *C. arabica - C. kahawae* samples relative to the application of different normalization factors. Main statistics given by the statistic test Kruskal-Wallis on thr1 expression for *C. arabica - C. kahawae* samples, comparing the normalization factors followed

https://doi.org/10.1371/journal.pone.0150651.s008

Table A4.4 - Statistical analysis of *thr1* expression in *C. kahawae* samples relative to the application of different normalization factors. Test statistics given by the Kruskal-Wallis statistic on the expression for *C. kahawae* samples, comparing the normalization factors followed

https://doi.org/10.1371/journal.pone.0150651.s009

Table A4.5 - Statistical analysis of *cat2* expression in *C. arabica - C. kahawae* samples relative to the application of different normalization factors. Test statistics given by the Kruskal-Wallis statistic test on *cat2* expression for *C. arabica-C. kahawae* samples, comparing the normalization factors followed

https://doi.org/10.1371/journal.pone.0150651.s010

Table A4.6- Statistical analysis of *cat2* expression in *C. kahawae* samples relative to the application of different normalization factors. Test statistics given by the Kruskal-Wallis statistic test on *cat2* expression for *C. kahawae* samples, comparing the normalization factors followed

https://doi.org/10.1371/journal.pone.0150651.s011

Table A4.7 - Statistical analysis of *thr1* expression between different time points. Test statistics given by the Mann-Whitney test on thr1 expression, comparing time points for *C.arabica - C.kahawae* and lifecycle stages for *C. kahawae*

https://doi.org/10.1371/journal.pone.0150651.s012

Table A4.8 Statistical analysis of *cat2* expression between different time points. Test statistics given by the Mann-Whitney test on *cat2* expression, comparing time points for *C.arabica - C.kahawae* and lifecycle stages for *C. kahawae*

https://doi.org/10.1371/journal.pone.0150651.s013